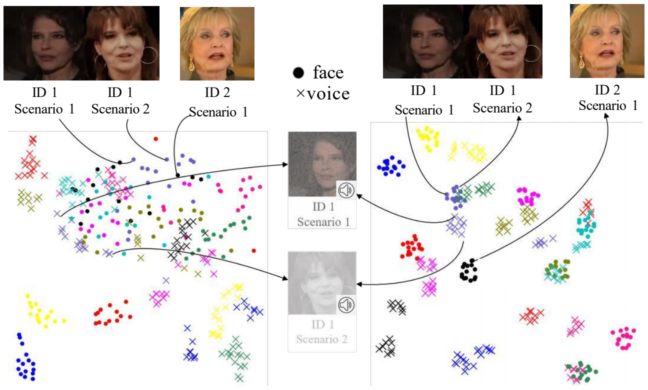
Public-private Attributes-based Variational Adversarial Network for Audio-Visual Cross-Modal Matching

Aihua Zheng, Fan Yuan†, Haichuan Zhang†, Jiaxiang Wang, Chao Tang, and Chenglong Li*

***Abstract*—Existing audio-visual cross-modal matching methods focus on mitigating cross-modal heterogeneity but ignore the impact of intra-class discrepancy of the same identity in different scenarios, which might greatly limit the matching performance. To simultaneously handle both problems of intra-class discrepancy and cross-modal heterogeneity, we propose a novel public-private attributes-based variational adversarial network (P^2VANet), which captures the consistency within and between classes, for audio-visual cross-modal matching. In particular, P^2VANet first uses a variational auto-encoder, which captures the inherent global information in diverse scenarios from the hidden variable through reconstruction, to reduce the intra-class discrepancy. Then it integrates a public attributes guidance module to capture the consistency of audio and visual by supervision of the common high-level semantic information to mitigate cross-modal heterogeneity. In addition, P^2VANet designs private attributes embedding module to enhance the discriminative features inherent in each class to decrease inter-class similarity. Extensive experiments on audio-visual cross-modal matching demonstrate the effectiveness of the proposed approach compared with the state-of-the-art methods.**

***Index Terms*—Audio-visual cross-modal matching, variational adversarial learning, public-private attributes, metric learning**



(a) Features extracted by traditional CNN after adversarial learning

(b) Features extracted by P^2VANet

Fig. 1. Visualization of the learned embedding of 10 identities from the training set in VoxCeleb [9] by t-SNE. Different colors indicate different identities.

I. Introduction

Existing cognitive and neuroscientific research has shown that humans can 'see voices' or 'hear faces'. Bruce and Young proposed that we can recall a personal identity through associative memory (e.g. face, voice, and other modalities) which suggests that we can associate the face of a particular person with the voice information and vice versa [1]. This task is named audio-visual learning (AVL). AVL includes audio-visual matching [2], audio-visual localization [3], audio-visual speech separation [4],audio-visual recognition [5],[6] and audio-visual Generation[7], [8]. It has crucial potential applications in criminal investigation, identity recognition, and film entertainment, to name a few. In certain scenarios, relying only on single-modal information is highly susceptible to interference, making the issue challenging to address.

The critical issue in AVL is to explore the essential connection between audio and visual which has attracted a great deal of interest. However, there is an inherent cross-modal heterogeneity between audio and visual modalities, which brings a great challenge in learning the cross-modal association.

In recent years, there has been an increasing number of work along this line. SVHF-Net [2] first proposed to view the matching task as a binary classification problem, and achieved great performance that outperforms the human benchmark by designing a two-stream network. However, the intrinsic connection between audio and visual is not fully explored. Wen *et al.* [10] proposed a disjoint mapping network (DIMNets) to learn a public representation by considering more covariates between audio and visual. However, the network fails to adequately compensate for modal heterogeneity by relying on covariates alone. Zheng *et al.* [11] proposed a novel adversarial metric learning (AML) model to overcome the heterogeneous issue by learning a modality-independent embedding. However, the network handles difficult samples and simple samples in the same way. Wen *et al.* [12] proposed to better explore the hard but valuable samples by designing a dynamic reweighting scheme. However, existing methods

†: F. Yuan and H. Zhang contribute equally to the article.

This research is supported in part by the National Natural Science Foundation of China (62372003), Natural Science Foundation of Anhui Province (No. 2308085Y40), and the University Synergy Innovation Program of Anhui Province (GXXT-2022-036).

A. Zheng, H. Zhang and C. Li are with Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, the Information Materials and Intelligent Sensing Laboratory of Anhui Province, School of Artificial Intelligence, Anhui University, Hefei, 230601, China (e-mail: ahzheng214@foxmail.com; zhc2000@foxmail.com; lcl1314@foxmail.com).

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

C. Tang is with the Department of Computer Science and Technology, Hefei University, Hefei, 230601, China (e-mail: tangchao77@sina.com)

mainly devote to implementing AVL by mitigating cross-modal heterogeneity through modal alignment, but ignore the large intra-class discrepancy across diverse scenarios, as shown in Fig. 1 (a). Due to the large changes in the scenarios, the face images of the same ID (such as ID1 in Scenario 1 and Scenario 2) present large discrepancies in both visual appearance and the feature distribution extracted by traditional CNN. This intra-class discrepancy brings a crucial challenge for audio-visual matching.

As a form of deeply generative model, Variational auto-encoder (VAE) is a generative test network structure based on variational Bayesian inference [13]. It uses two neural networks to establish two probability distribution models. One is used for variational inference of the original input data to generate the variational probability distribution of the hidden variable, which is called the inference network. The other is to reconstruct the original input data based on the generated hidden variables.

In contrast to traditional auto-encoder, VAE treat the encoder's output as a random variable and introduce an additional probability distribution to represent this variable, thus incorporating stochasticity. Generated samples are compelled to conform to specific distributions (e.g., Gaussian distribution), and constraints are applied to the generated data for each sample during the decoding process. These constraints help reduce differences between samples, thereby diminishing intra-class discrepancy. Moreover, VAE learns a more compact representation within the encoder's hidden layer, reducing sample differences by decreasing redundant information. By compressing the feature representations of intra-class samples into a more compact space, intra-class variability can be effectively reduced. Based on these observations, we propose utilizing VAE to constrain generation, minimizing redundant information within samples and suppressing intra-class discrepancy across scenes. As shown in Fig. 1 (b), after implementing VAE, samples from different scenes but belonging to the same identity exhibit a more clustered tendency, significantly reducing intra-class discrepancy. Note that the audio features appear much milder intra-class discrepancy compared with the visual face data, as shown in Fig. 1. However, the unavoidable noise in the audio may disrupt the intra-class compactness within the audio data, shown as the voice features of ID1 in scenario 1 and scenario 2. Therefore, we simulate the prior distribution of clean audio by using encoder downscale to eliminate noise and constrain the distribution between the original audio and clean audio [14]. The audio VAE learns from the corrupted input, which provides noisy input to the encoder network, and then compares the decoder's reconstructed data to the original input, teaching the network how to denoise the input.

Adversarial learning (AL) has been evidenced as an effective way to mitigate audio-visual cross-modal heterogeneity [15], [16] by the competition between the modality-independent feature generator and the modality discriminator. However, existing adversarial learning methods [11], [17], [18] attempt to directly capture the relationship between the low-level visual and audio features, while ignoring their high-level semantic consistency. This remains a large semantic gap between audio and visual data. Therefore, we propose to strengthen the semantic connection by the supervision of public/common attributes, such as nationality and gender between the visual and the associated audio data into the adversarial learning framework. We supervise the high-level semantic information through the guidance of the public attributes classifiers, which further enforce the cross-modal sample distributions. Based on this, the public attributes can further mitigate audio-visual cross-modal heterogeneity, shown as the visual and audio features in Fig. 1, samples from different modalities are aggregated together after introducing the public attributes guidance.

Most existing AVL methods mainly employ the cross-entropy loss supervised by identity labels into constraining the recognition/classification. However, it is still difficult to distinguish between these different classes with huge inter-class similarity challenge from a global perspective, shown as the visual features of ID1 in scenario 2 and ID2 in scenario 1 in Fig. 1. Private attributes of visual and audio, such as nose size, pitch level, etc., contain rich semantic information and enable critical discriminative fine-grained features to distinguish similar classes. Therefore, we propose to embed these fine-grained discriminative features to further reduce the inter-class similarity. Specifically, we incorporate these private attributes into the hidden variable via cross-transformer [19] to obtain both global and fine-grained information.

To our best knowledge, it is the first work to investigate to utilize variational learning to address intra-class variability and leverage public attributes and private attributes to mitigate modality heterogeneity and reduce inter-class similarity, named as P^2**VANet** in this paper. The contribution of this paper can be summarized as:

- To narrow intra-class discrepancy and simultaneously remove noise from audio, we propose to compress the feature representations of intra-class samples into a more compact space.
- To further mitigate cross-modal heterogeneity, we propose to integrate the public attributes into the adversarial learning framework to supervise the high-level semantic information.
- To reduce inter-class similarity, we propose to utilize private attributes to acquire critical discriminative fine-grained features with rich semantic information to distinguish different classes.
- Extensive experiments demonstrate the effectiveness of the proposed framework for audio-visual cross-modal matching while handling the three issues in audio-visual matching in comparison with the state-of-art methods.

II. Related works

A. Audio-Visual Matching

The cognitive studies have demonstrated that it is possible to discover the association between voices and faces, and the machine learning community is showing increasing interest in studying such association [2]. Based on this, many works performed speaker recognition by jointly observing audio and visual signals. By assuming that voices and faces are

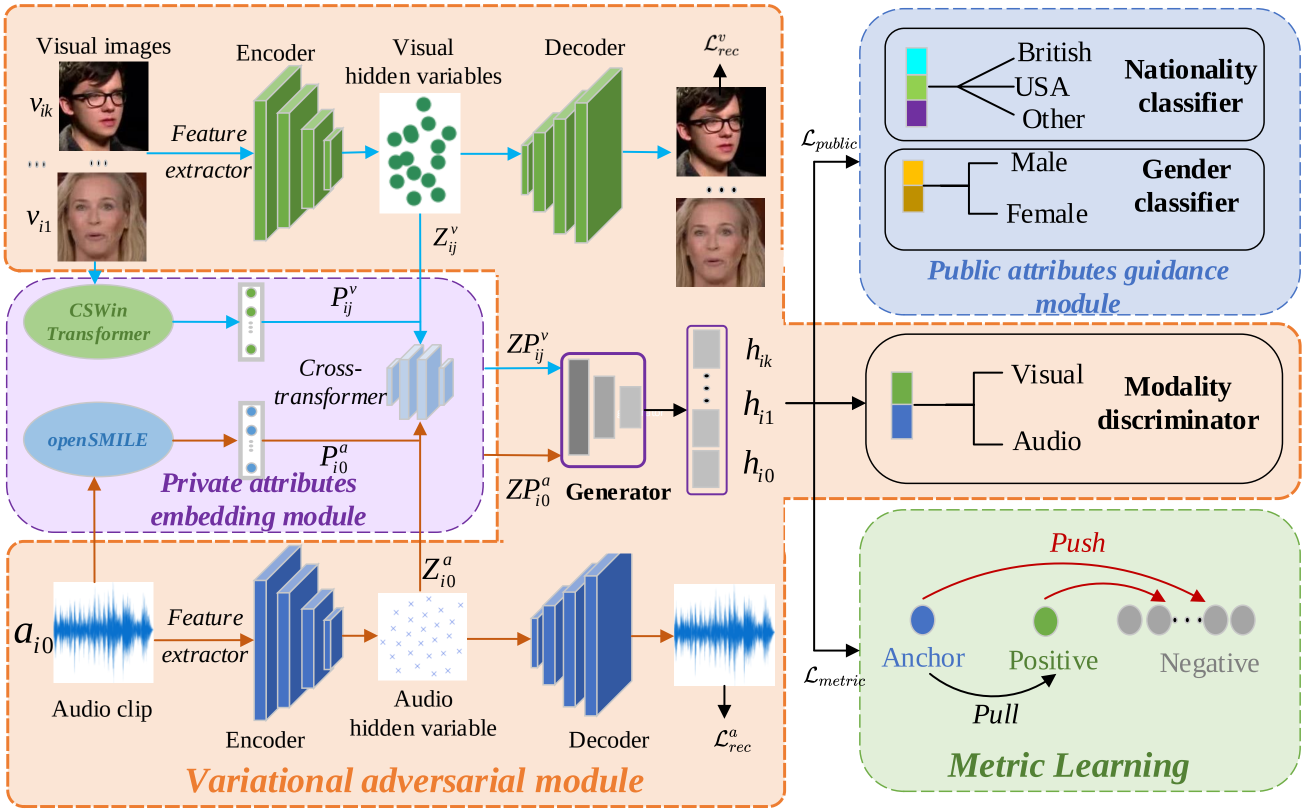


Fig. 2. Schematic overview of our proposed P^2VANet model.
Visual images and audio clip are fed into the visual encoder and the audio encoder to obtain visual hidden variables Z^v_{ij} and audio hidden variable Z^a_{i0}, respectively. After obtaining the visual private attributes P^v_{ij} and audio private attributes P^a_{i0} by CSWin-transformer and openSMILE respectively, we propose to learn the hidden variable with private attributes ZP^v_{ij}, ZP^a_{i0} via cross-transformer. After that, we propose to learn the modality-independent feature embeddings $\{h_{i0}, h_{i1}..., h_{ik}\}$ by adversarial learning. In addition, we propose to strengthen the semantic connection by the supervision of public attributes to further mitigate cross-modal heterogeneity. Meanwhile, we propose to learn robust feature embeddings for similarity measure via metric learning.

implicitly captured from a speaker, these works learned a public embedding through the joint representation of voices and faces, maximizing their similarity if they belong to the same speaker [18].

Nagrani *et al.* [2] first viewed the matching task as a classification problem by proposing the SVHF-Net which consists of a static network, a dynamic network, and an N-way network. However, the static and dynamic network only handle a certain number of images and audio tracks. Nagrani *et al.* [1] extended the audio-visual matching task to audio-visual verification and retrieval, proposing a joint optimization embedding network that incorporates both curriculum learning and contrastive loss. Visual and auditory features are feded into a shared feature space to learn a common representation that captures information from both modalities. However, it has not fully learned the intrinsic correlation between audio and visual modalities. Wen *et al.* [10] proposed to use covariates as restriction to learn the correlation between audio and visual modalities. However, it increased a large number of parameters into the network while introducing covariates between every audio-visual pair. Nawaz *et al.* [20] proposed a single-stream network with shared parameter feature representation between audio and visual information to handle the difficult training with exceeded parameters. However, all of these methods ignore the intrinsic heterogeneity between the two modalities. Based on this, Zheng *et al.* [11] and Wang *et al.* [16] proposed to use adversarial learning to generate modality-independent feature representation to bridge the modality gap, followed by the metric learning to further learn a robust similarity measure for cross-modality. Yu *et al.* [21] and Saeed *et al.* [22] suggested to obtain shared and specific features for the audio-visual modality through orthogonal decoupling, yet still ignored the attribute covariates in modalities.

B. Variational Auto-Encoder

VAE is a deeply hidden space generation model, which has been widely used in image generation. VAE employs neural networks to encode and decode data, and generates reconstructed data that resembles the original input, has been successfully employed in both computer vision and machine learning tasks [14], [23]. However, the traditional VAE has many hypothetical preconditions and constraints, and the generated images are often blurry, the expressive power of complex models is poor. In order to solve this problem, researchers have also proposed many variants of VAE.

Conditional variational auto-encoder (CVAE) [24] is proposed for target generation of sample data of specific categories. The CVAE with category information tags has also changed from the traditional unsupervised model to the semi-supervised model. Variational fairness auto-encoder (VFAE) [25] is used to separate the noise factor from the hidden variable, so that the model can learn the characteristic representation of some non-degeneration factors more clearly. Important weighted auto-encoder (IWAE) [26] is a new model proposed to solve the poor generalization ability of traditional VAE. IWAE hoped to increase the number of samples of the hidden variable corresponding to the original sample to make the lower bound of the variation more compact and closer to the true log-ikelihood, thereby improving the adaptability of the generation network to different distributions. Variational loss auto-encoder (VLAE) [27] aimed to use the autoregressive neural network model's powerful generation ability to transform VAE.

Nouveau VAE (NVAE) [28] concluded that most of the research in VAE has focused on statistical challenges and explored an orthogonal direction of designing specific neural architectures for hierarchical VAE. NVAE is a hierarchical VAE that generates images using deep separable convolution and batch normalization techniques .

C. Attributes Guided Embedding

In previous research, multiple studies have focused on implementing attribute-guided embedding to investigate the impact of incorporating attribute supervision in guiding discriminative tasks. Wen *et al.* [10] learned shared representations of different modalities by treating attributes as common covariates of different modalities. These shared representations can then be used to find correspondences between modalities. Li *et al.* [29] introduced an attribute and state guided structural embedding network for the vehicle re-identification task. This network achieves discriminative feature learning through attribute-based enhancement and state-based weakening. Chai *et al.* [30] proposed enhancing crucial image regions using relevant attributes such as gender, age, and clothing features, allowing this module to focus more accurately on the pedestrian's body and better distinguish the target from the background for identification. Cheng *et al.* [31] proposed a visual-textual baseline for pedestrian attribute recognition, which encodes pedestrian images and attribute annotations as visual and textual features, respectively. The model interacts with cross-modal information and explores textual-semantic associations in the attribute annotations using a pre-trained text encoder. Chen *et al.* [32] proposed an identity-aware contrast distillation loss to guide the student network to effectively learn the interrelationships among multi-attribute samples. These studies have shown that attribute-guided embedding can significantly increase the distance of inter-class samples and decrease the distance of intra-class samples. In contrast to these approaches, our method takes into account not only public attributes like gender and nationality, but also modal private attributes. Additionally, our approach incorporates private attributes into the embedding through attention interactions, effectively leveraging both public and private attributes to supervise the network.

III. Methods

To relieve the intra-class discrepancy, inter-class similarity, and the cross-modal heterogeneity in audio-visual matching, we propose a variational adversarial network for audio-visual matching based on public-private attributes which is called P^2**VANet** in this paper, as shown in Fig. 2.

We utilize OpenSmile and SwinTransformer as our audio and visual feature extractors, respectively. Subsequently, our focus is on designing the extraction of hidden variables for audio-visual modalities and their interaction with attribute features. In order to alleviate the effects of intra-class discrepancy caused by the large changes of the scenarios, we propose to use variational learning to learn the inherent global information among diverse scenarios. Specifically, we reconstruct from the hidden variable to preserve the inherent global information, which presents less variation across scenarios compared to average features. To reduce the inter-class similarity, we use transformer-based private attributes to enrich the discriminative fine-grained information in the hidden variable. Specifically, we embed the pre-extracted private attributes into the hidden variable of the corresponding instances via cross-transformer [19] to focus on both global and fine-grained information. To mitigate the cross-modal heterogeneity, we propose to use public attributes to supervise the high-level semantic information. Specifically, we design a gender classifier and a nationality classifier for each instance separately to enhance the consistency of cross-modal data in high-level semantic information.

A. Problem formulation and Baseline

For the sake of generality and simplicity, taking the voice to face (V-F) task in audio-visual matching as an example, given an audio clip a_{i0} as a query and a gallery consisting of k visual images, the gallery include a positive sample \$$v_{i1}$ and $k-1$ negative samples $\{v_{i2}, ..., v_{ik}\}$, audio-visual matching aims to discover the corresponding facial images from the gallery for querying speech audio. Specifically, the gallery size $k \geq 2$ denotes a matching task, where k = 2 denotes a binary matching case, otherwise it denotes a multi-way matching case. We extract audio feature f^a_{i0} and visual features $\{f^v_{i1}, f^v_{i2}, ..., f^v_{ik}\}$ separately. The F-V challenge can be defined in the same way. The entire network consists of five convolutional layers. Our baseline approach comprises an audio-visual feature extractor, an adversarial learning generator, and a discriminator. This setup facilitates the learning of pattern-independent feature representations by engaging in a min-max game between the feature generator and the pattern discriminator.

B. Variational Adversarial Module

Variational learning. In order to alleviate the intra-class discrepancy, we propose to use the variational auto-encoder to constrain generation, reducing redundant information within samples across different scenes.

To simplify the illustration of the equations, we first elaborate the operation in the audio modality in the following discussion. Variational auto-encoder consists of an encoder and a decoder. The encoder takes audio feature f^a_{i0} as input and employs variational inference techniques to learn the distribution of the original data, generating hidden variables to represent the underlying features. The decoder extracts the hidden variables from the latent space and maps them back to the original data space. It reconstructs the input data by receiving and sampling the hidden variables representation generated from the encoder, thus helping the network better focus on intra-class information. Specifically, we obtain the mean μ^a_{i0} and variance σ^a_{i0} through the encoding process and assume that the posterior distribution of the samples follows a Gaussian distribution. The hidden variable after reparameterization [13] we sampled is:

$$Z^a_{i0} = \mu^a_{i0} + e * \sqrt{\sigma^a_{i0}}, \tag{1}$$

The ultimate goal of the variational auto-encoder in our method is to obtain the hidden variable which contains inherent global information by constraining the reconstructed data R^a_{i0} and the input data f^a_{i0}. The reconstruction loss $\mathcal{L}^a_{rec}$ minimized during the training has the following form:

$$\mathcal{L}^a_{rec} = (R^a_{i0} - f^a_{i0})^2 + KL(q(Z^a_{i0}|f^a_{i0})||p(Z^a_{i0})), \tag{2}$$

$$KL(p(x)||q(x)) = \int_x p(x) log \frac{p(x)}{q(x)}, \tag{3}$$

where $q(Z_{i0}|f_{i0})$ denotes the posterior distribution of the potential variable Z_{i0} corresponding to the input data f_{i0}. KL denotes the Kullback-Leibler divergence, which measures the similarity between the posterior and the generative distributions. Similarly, the reconstruction loss of visual modality can be defined as:

$$\mathcal{L}^v_{rec} = \sum_{j=1}^{k} [(R^v_{ij} - f^v_{ij})^2 + KL(q(Z^v_{ij}|f^v_{ij})||p(Z^v_{ij}))], \tag{4}$$

where R^v_{ij}, f^v_{ij}, and Z^v_{ij} denote the reconstructed data, input data, and hidden variables of the visual modality, respectively. We retain the global information by resampling the hidden variable (Eq. (1)) and reconstructing (Eq. (2)), (Eq. (4)). The reconstruction target of the network is forced to be close to the original features in order to retain as much global information as possible and reduce the loss of information.

Liu et al. [34] and Deng et al. [35] demonstrate that variational learning can effectively reduce intra-class discrepancy. To better illustrate the effectiveness of variational learning, we employ 2D t-SNE [33] to visualize the audio and visual feature distribution before and after variational learning, as shown in Fig. 3 and Fig. 4 respectively. Variational learning enables the network to focus on a wider area and to have more information contained in the learned features. As shown in Fig. 3, the audio features of ID1 in scenario 1, scenario 2 and scenario 3 in Fig. 3 become closer after variational learning. Fig. 4 demonstrates the consistent visualization. The facial images of the same individual with varying poses are heavily dispersive before variational learning, especially in Scenario 1. By contrast, variational learning can effectively bring facial features with varying poses into a compact space. Fig. 3 and Fig. 4 demonstrate that variational learning can alleviate the intra-class discrepancy issue between different scenarios.

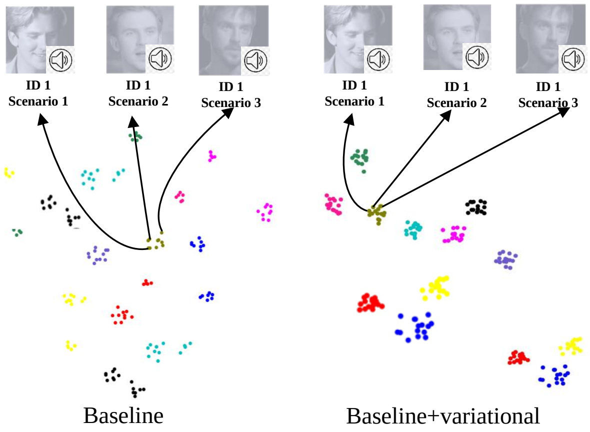


Fig. 3. 2D t-SNE [33] visualization of the **audio** feature representation before and after the variational learning. Different colors represent different identities.

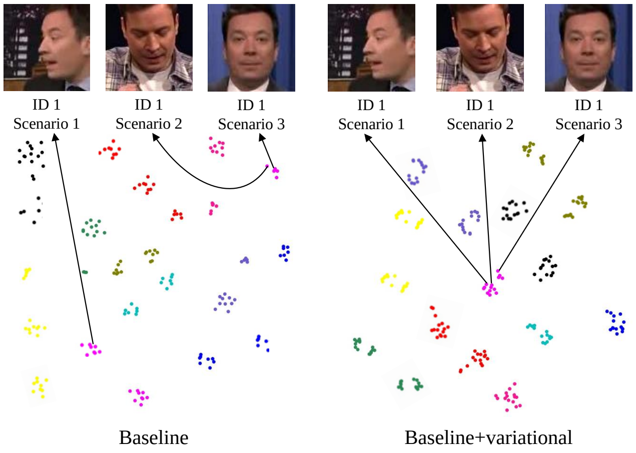


Fig. 4. 2D t-SNE [33] visualization of the **visual** feature representation with diverse poses before and after the variational learning. Different colors represent different identities.

Adversarial learning. To mitigate the cross-modal heterogeneity, we introduce using adversarial learning to learn modality-independent feature embeddings between audio and visual.

In our methods, the generator G is constructed using standard MLP to learn two mapping functions Ga and Gv, which beats with the modality discriminator D through min-max game. The mapping functions Ga takes audio feature f^a_{i0} and Gv takes visual features $\{f^v_{i1}, f^v_{i2}, ..., f^v_{ik}\}$ as input, and aims to generate modality-independent feature embeddings $\{h_{i0}, h_{i1}, ..., h_{ik}\}$. The modality discriminator D is designed as a modality classifier with parameter θ_D, which is consists of a fully connected network to distinguish the original modality of the modality-independent feature embeddings $\{h_{i0}, h_{i1}, ..., h_{ik}\}$ obtained from the generator. The discriminator loss $\mathcal{L}_D$ is trained by minimizing:

$$\mathcal{L}_D = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} \sum_{j=0}^{k} y_{ij} \log D(h_{ij}; \theta_D), \tag{5}$$

where y_{ij} represents the modality label of the j-th sample in the i-th instance group, $D(h_{ij}; \theta_D)$ is the modal possibility of the generator output, M_{train} represents the number of training instance groups. We employed Eq. (5) for the modal discrimination task to achieve modal equalization.

We observe that the blue points became farther after variational learning, likely due to the fact that the same identity may not convey the same content in different audio clips, resulting in variations after embedding. While we theoretically anticipate that mining more features through variance learning should tighten the inter-class distance, there are cases where variance learning fails to bring intra-class samples closer. Therefore, we incorporate attributes supervision to further address these instances.

C. Private Attributes Embedding Module

To further distinguish different classes with huge inter-class similarity, we propose to utilize fine-grained discriminative features with rich semantic information to distinguish similar classes. For visual information, we first extract the 40-dimensional face attributes mentioned in CelebA [36] from VGGFace [37] via transformer. Considering that the face images in our task come from multiple scenarios, the values of these attributes change under different scenarios. Therefore, we selected the 12 invariant face attributes (e.g. nose size) as our visual private attributes. Specifically, we use CSWin-transformer [38] to extract visual private attributes which are represented as a one-hot vector. For audio information, we use the openSMILE [39] to obtain the 384-dimensional features [40] as our audio private attributes. Segment-level statistical heuristic features are used to expand the differences between audio segments among different identities. These features include but are not limited to, voice prosody features, voice quality features, zero-crossing rate, etc.

To simplify the illustration of the equations, we only elaborate the operation in the audio modality in the following discussion and the visual modality can be implemented in the same manner. As shown in the left part of Fig. 5, after obtaining the hidden variable Z^a_{i0} and the audio private attributes P^a_{i0}, we employ the cross-transformer [19] to embed the audio private attributes into the hidden variable. The hidden variable and audio private attributes are first delivered to an switchable-norm (SN), which is defined as:

$$Z^a_{SN} = SN(Z^a_{i0}), \tag{6}$$

$$P^a_{SN} = SN(P^a_{i0}), \tag{7}$$

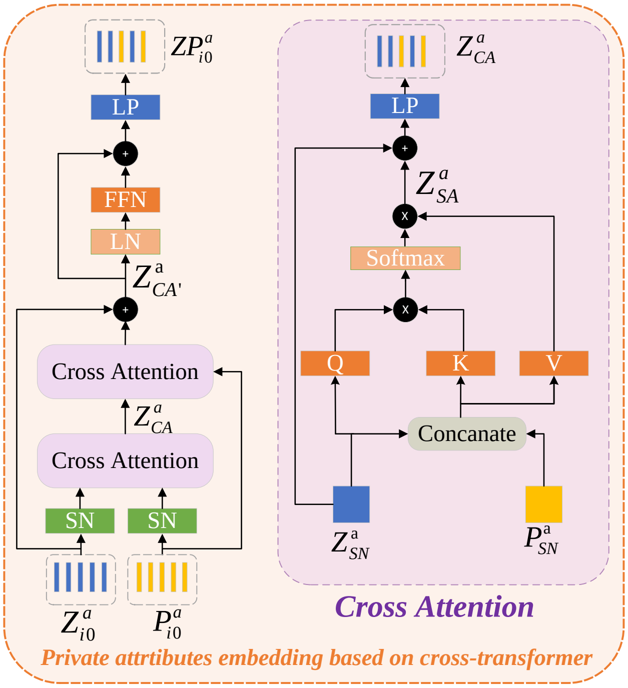


Fig. 5. Architecture of the cross-transformer for embedding the private attributes into the audio hidden variable.

where Z^a_{SN}, P^a_{SN} are the normalized features. In contrast to the use of layer-norm (LN), $SN(\cdot)$ determines the appropriate normalization operation for each normalization layer by dynamically adjusting the weights of the various normalization methods, allowing the network to determine which normalization method to adopt. Normalized features are sent to the cross attention block to integrate the two features. This can be formulated as:

$$Z^a_{CA} = CA(Z^a_{SN}, P^a_{SN}), \tag{8}$$

where $CA(\cdot)$ is the cross attention module, as shown in the right part of Fig. 5.

The cross-attention module is a modified multi-headed attention module that captures features from the private attributes that contribute to the hidden variable. We use the hidden variable as a query that interacts with the private attributes through attention. We use the normalized features Z^a_{SN} as the query (Q), and connect them to the features from the other branch P^a_{SN} to serve as the key (K) and value (V). The connection between the two features is :

$$Z^a_{SA} = softmax(\frac{QK^T}{\sqrt{d/h}})V, \tag{9}$$

where d is the dimension of aligned features Z^a_{SN} and h is the number of heads in the cross attention model. The output of the cross attention module Z^a_{CA} can be defined as:

$$Z^a_{CA} = LP(Z^a_{SA} + Z^a_{SN}), \tag{10}$$

where $LP(\cdot)$ represents linearly projected. In order to more fully embed the information in private attributes, we propose a multi-stage integration approach. Specifically, the output of

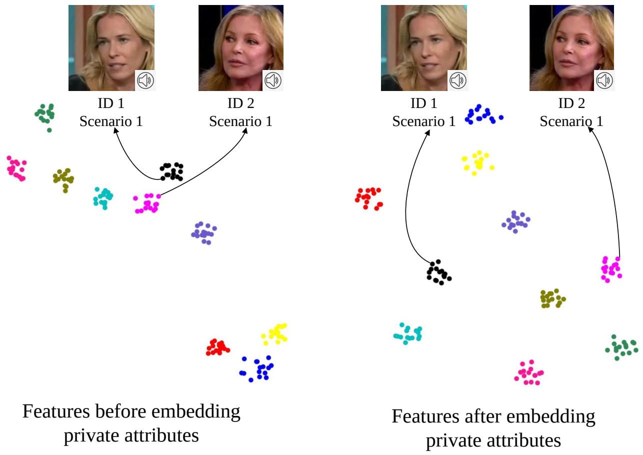


Fig. 6. The visualization of the audio feature representation before and after embedding with private attributes. Different colors represent different identities.

the cross attention block Z^a_{CA} and the private attributes P^a_{i0} are fed into the cross attention block again as inputs and are summed with the hidden variable Z^a_{i0} to obtain the new output $Z^a_{CA'}$:

$$Z^a_{CA'} = CA(Z^a_{CA}, P^a_{i0}) + Z^a_{i0}, \tag{11}$$

The hidden variable after integration of private attributes ZP^a_{i0} is defined as:

$$ZP^a_{i0} = LP(FFN(LN(Z^a_{CA'})) + Z^a_{CA'}), \tag{12}$$

where $LN(\cdot)$ is the layer-norm, $FFN(\cdot)$ denotes a feedforward network.

Previous works [41], [29] have demonstrated that private attributes have an effective bootstrapping effect on the identification task, which can effectively make the inter-class samples more distinguishable and reduce the intra-class sample distance. As shown in Fig. 6, features from different identities become farther after integrating private attributes. Also, the sample distances within classes become more compact, compared to the right column of Fig. 3. Private attributes provide discriminative fine-grained features for each sample, which increase the difference between classes in terms of local information and mitigate inter-class similarity to some extent.

D. Public Attributes guidance Module

Adversarial learning has been shown to compensate for cross-modal heterogeneity [11], [18], [42], [17]. However, it ignores the consistency of cross-modal data in high-level semantic information. We propose to strengthen the high-level semantic connection by the supervision of public attributes. We supervise the high-level semantic information through the guidance of the public attribute classifiers. Specifically, we use cross-entropy loss[43] to provide a loss function for each public attribute. The loss functions for the gender $\mathcal{L}_{gender}$ and the nationality $\mathcal{L}_{nation}$ are defined respectively as:

$$\mathcal{L}_{gender} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} \sum_{j=0}^{k} g_{ij} \log C_G(h_{ij}; \theta_{C_G}), \tag{13}$$

$$\mathcal{L}_{nation} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} \sum_{j=0}^{k} n_{ij} \log C_N(h_{ij}; \theta_{C_N}), \tag{14}$$

where g_{ij} and n_{ij} represent the gender label and nation label of the j-th sample in the i-th instance group respectively. $C_G(h_{ij}; \theta_{C_G})$ and $C_N(h_{ij}; \theta_{C_N})$ represent the gender and nation possibility of the generator output, M_{train} represents the number of training instance groups. The loss function of the public privates is defined as the sum of the both, i.e., $\mathcal{L}_{public} = \mathcal{L}_{gender} + \mathcal{L}_{nation}$. Every classifier is all simple multi-class logistic regression classifier including a single softmax layer. The public attributes can further mitigate audio-visual cross-modal heterogeneity by enforcing the cross-modal sample distributions.

E. Metric Learning

In order to further reduce the effects of intra-class discrepancy and inter-class similarity, we use metric learning to learn a robust similarity metric for audio-visual matching. Similar samples should have closer distances and dissimilar samples should have farther distances. Specifically, we follow lifted structured loss [44] to constrain all positive and negative pairs of samples in the training set. The definition of the optimization objective is:

$$\mathcal{L}_{metric} = \frac{1}{2M_{train}} \sum_{i=1}^{M_{train}} max((0, \mathcal{J}_i))^2, \tag{15}$$

$$\mathcal{J}_i = log(\sum_{j\in[2,k]} e^{\theta - d_{i0,ij}} + \sum_{q\in[2,k]} e^{\theta - d_{i1,iq}}) + d_{i0,i1}, \tag{16}$$

where $d_{i0,i1}$ measures the Euclidean distance between anchor h_{i0} and positive sample h_{i1}, $d_{i0,ij}$ represents the Euclidean distance between anchor h_{i0} and negative samples h_{ij}, $d_{i1,iq}$ measure the Euclidean distance between positive sample h_{i1} and negative samples h_{iq}. $\mathcal{J}_i$ is the distance used to measure the similarity of the positive pair and dissimilarity between negative samples and the positive pair.

F. Training

During training, for every epoch, we encode audio and visual data separately to obtain the corresponding mean μ and variance σ that conform to the Gaussian distribution. The hidden variable is sampled from $\mathcal{N}(\mu, \sigma)$. Then we integrate the hidden variable and the pre-extracted private attributes to form the hidden variable with fine-grained information. The hidden variable with fine-grained information is spliced and sent to the modality discriminator, public attributes guidance module, and metric learning module. After public-private attributes-based variational adversarial learning, we use the distance between a given audio clip and each visual sample for matching evaluation. The matching classifier is designed as a fully connected network. We concatenate the modality-independent feature embeddings and feed them into the matching classifier, using the cross-entropy loss[43] as our classification loss $\mathcal{L}_{cls}$:

$$\mathcal{L}_{cls} = -\frac{1}{M_{train}} \sum_{i=1}^{M_{train}} l_i \log C_M(\{h_{i0}, h_{i1}, ..., h_{ik}\}; \sigma_{C_M}), \tag{17}$$

TABLE I
THE QUALITATIVE RESULTS OF MATCHING AND VERIFICATION TASKS ON VOXCELEB. BINARY DENOTES THE 1:2 MATCHING WHILE MULTI-WAY DENOTES THE $1 : k$ $(k = 10)$ MATCHING. '×' INDICATES 'NOT CAPABLE' AND '-' INDICATES 'NOT AVAILABLE'.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (AUC)	
		V-F	F-V	V-F	F-V	V-F	F-V
SVHF-Net [2]	CVPR2018	81.0	79.5	34.5	×	–	–
DIMNet [10]	ICLR2019	81.3	81.9	38.4	36.2	81.0	81.2
Wang's [45]	ACM2020	83.4	84.2	39.7	36.4	82.6	82.9
Wen's [12]	CVPR2021	87.2	86.5	48.2	44.8	87.2	87.0
AML [11]	TMM2021	89.4	86.6	46.2	43.7	86.4	86.2
DSANet [16]	TMM2022	92.4	88.4	49.1	46.7	87.3	87.4
DCLR [21]	ICDM2022	86.8	87.5	48.3	46.2	86.7	86.9
SBNet [22]	ICASSP2023	82.4	82.4	-	-	82.5	82.6
P^2VANet	Ours	**93.1**	**90.4**	**50.6**	**48.1**	**88.5**	**88.7**

TABLE II
THE QUALITATIVE RESULTS OF MATCHING TASKS ON VOXCELEB2. BINARY DENOTES THE 1:2 MATCHING WHILE MULTI-WAY DENOTES THE $1 : k(k = 10)$ MATCHING. '×' INDICATES 'NOT CAPABLE' AND '-' INDICATES 'NO RESULTS'.

Methods	Venue	Binary (ACC)		Multi-way (ACC)		Verification (AUC)	
		V-F	F-V	V-F	F-V	V-F	F-V
SVHF-Net [2]	CVPR2018	68.7	67.9	×	×	–	–
DIMNet [10]	ICLR2019	68.5	69.0	–	–	–	–
AML [11]	TMM2021	80.2	81.4	41.2	40.7	80.6	78.4
P^2VANet	Ours	**87.3**	**85.2**	**46.2**	**45.1**	**84.9**	**82.1**

where C_M denotes the audio-visual matching classifier, which is used to calculate the probability that an audio clip belongs to each visual image. σ_{C_M} is the parameters of the classifier and l_i is the i-th label. Eq. (17) is used for the identity matching task to achieve identity prediction. Our model integrates variational learning, metric learning into adversarial learning to joint learn the modality-independent feature embeddings. We update our model by minimizing the total loss $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = \mathcal{L}_{metric} + \lambda_1 \mathcal{L}_{rec} + \lambda_2 \mathcal{L}_{cls} + \lambda_3 \mathcal{L}_{gender} + \lambda_4 \mathcal{L}_{gender} \tag{18}$$

where $\mathcal{L}_{rec} = \mathcal{L}^a_{rec} + \mathcal{L}^v_{rec}$. λ_1, λ_2, λ_3 and λ_4 are the hyper-parameters of the balance loss, which are empirically set to 1.5, 2, 1, 1.5 in the experiments.

IV. RESULTS

In this section, we will introduce the details of our experiment, including audio-visual matching and verification, and compare our results with existing methods.

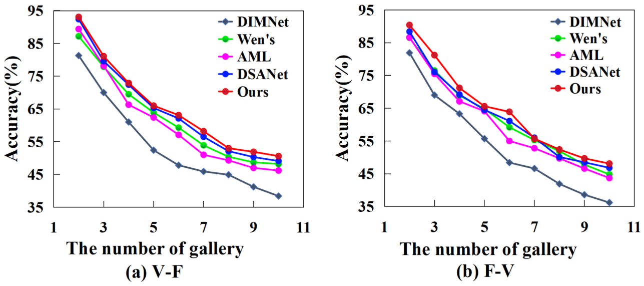
A. Datasets

VoxCeleb. Following the protocols in Wen *et al.* [12], we evaluate our method on benchmark VoxCeleb [9] and VGGFace [37] datasets. VoxCeleb [9] is an audio-visual dataset of speaker speech segments, which provides the audio data while VGGFace is a face dataset with corresponding face images. We use the face and voice segments of 1225 identities crossed in VoxCeleb and VGGFace for our task. Following the protocols in [2] [10], all identities with names beginning with "A" or "B" were retained for validation, while identities with names beginning with "C", "D" or "E" were retained for testing, and identities with names beginning with "F" - "Z" were used for training. This provides a good balance of male and female speakers.These three parts of the data do not overlap with each other.

VoxCeleb2. To further validate the effectiveness of our method, we also evaluate our method on Voxceleb2 [46] for the first time. VoxCeleb2 [46] is the successor to the Voxceleb [9] with larger quantities and more diversities. It has nearly double number of speakers and contains over one million speech segments from 6112 speakers, with approximately 16,100 hours of audio data, making it one of the largest publicly available speaker recognition datasets.

B. Implementation Details

1) Network architecture: The inputs are visual images with the shape of $224 * 224 * 3$ and audio clip spectrogram with $224 * 125 * 1$, respectively, which are normalized to [-1, 1], and the output is a 128-dimensional feature embedding. In the variational learning section, the input is a 128-dimensional feature after feature extractor, which is used to generate the hidden variable containing global information. The hidden variable is used for two functions: (1) as input to the decoder to reconstruct the original data, and (2) to integrate with private attributes and generate the modality-independent feature embeddings after adversarial learning. The modality-independent feature embeddings are utilized as the entry for the modality discriminator, the public attributes guidance module which are composed of several FC layers and the metric learning module.



(a) V-F

(b) F-V

Fig. 7. Matching results with varying number of k in $1 : k$ matching case in both V-F and F-V tasks.

2) Training parameters: The training process is divided into three stages: private attributes pre-extracted, variational adversarial learning and public attributes guidance. Variational learning, modality discriminator, gender classifier and nationality classifier with initial learning rates of $1*10^{-2}$, $5*10^{-3}$, $5*10^{-3}$, $5*10^{-2}$, respectively. The number of batch_size is 64. In the training phase, we use Adaptive Moment Estimation (Adam) with the momentum of 0.9 and weight decay of 0.0005 to fine-tune the network.

3) Competitors: We compare our method with state-of-the-art audio-visual matching methods including AML [11], Wen *et al.* [12], SVHF-Net [2], DIMNet [10], Wang *et al.* [45], DCLR [21], DSANet [16] and SBNet [22].

C. Comparison Results

Table I and Table II report the audio-visual matching and verification tasks in both V-F and F-V scenarios compared to the state-of-the-art methods on VoxCeleb [9] and redVoxceleb2 [46] respectively. V-F represents voice as the anchor while faces as the gallery, vise-versa for F-V. ACC means accuracy and AUC means Area Under Curve of the ROC (Receiver Operating Characteristic) curve. Clearly, our method beats the state-of-the-art methods in all the tasks and scenarios, which verifies the effectiveness of our method to suppress the intra-class discrepancy and mitigate the cross-modal heterogeneity. Compared to AML [11], our method introduces variational learning based on adversarial learning to reduce the intra-class differences caused by scene changes. Furthermore, we utilize public attributes as supervision to further mitigate modality heterogeneity, and employ private attributes to decrease inter-class similarity.

Although DIMNet [10] and DSANet [16] detach the discriminative feature attributes to enhance the face-voice association, the performance is still overshadowed. The main reason is that the unevenly distributed nationality and the limited gender information are not sufficient to constrain the association of audio and visual. By contrast, we propose to use them as guidance to supervise high-level semantic information. From binary to multi-way, as the number of candidates increases, the impact of the intra-class discrepancy between different scenarios on discrimination increases. Therefore, the matching accuracy decrease in both V-F and F-V scenarios.

TABLE III
ABLATION STUDY IN THE BINARY MATCHING TASK. 'PRI' AND 'PUB' DENOTE PRIVATE ATTRIBUTES AND PUBLIC ATTRIBUTES RESPECTIVELY.

Methods	V-F	F-V
Baseline	89.4	86.6
+ VAE	90.8	88.2
+ VAE + pri	92.7	89.6
+ VAE + nation	91.0	89.1
+ VAE + gender	91.4	89.3
+ VAE + pub	92.1	89.9
+ VAE + pri + pub (ours)	**93.1**	**90.4**

Even though, P^2VANet significantly outperforms the state-of-the-art methods by a large margin, which validates the robustness of our method while handling more challenging cases. The performance on VoxCeleb2 [46] generally declines since the large distribution diversity between the training and testing data, and it contains a larger number of speakers and a richer variety of audio samples. However, our method still achieves superior performance, indicating the effectiveness of the proposed method. Fig. 7 shows the detailed results of our method compared with other methods on VoxCeleb [9] in the multi-way case with the number of instances in the gallery ranging from 2 to 10. As the number of candidate instances increases, the likelihood of the same identity coming from different scenarios increases. Therefore, the matching challenge consequentially increases, which results in the performance decreasing. However, our approach consistently delivers higher performance and less reduction than the competitors, which suggests the robustness of our model.

D. Ablation Study

To verify the three components in the methodology of this paper, VAE, public attributes guidance, and private attributes embedding, we evaluate four ablated variants as shown in Table III and Table IV. All the VAE, public attributes guidance, and private attributes embedding make a positive contribution to our method by consistent improvement in both V-F and F-V cases in binary and multi-way scenarios. Note that in Voxceleb [9], the public attributes include gender and nationality, while in Voxceleb2 [46], only gender is included. First, both private attributes and public attributes play a positive role in the model. By introducing VAE into the baseline, our model achieves clear improvements on both Voxceleb [9] and Voxceleb2 [46], which evidences the effectiveness of VAE while reducing intra-class variability, especially on more challenging Voxceleb2 [46] dataset.

By further introducing the private and public attributes respectively, our model also achieves significant improvements. This verifies the contribution of the private and public attributes by weakly supervising semantic consistency across modalities and strengthening the correlated features respectively. It can be observed that the improvement on Voxceleb [9] is greater than that on Voxceleb2 [46]. This is because as the data volume increases, the number of speakers with the same attribute information also increases, which increases

TABLE IV
ABLATION STUDY IN THE BINARY MATCHING TASK ON VOXCELEB AND VOXCELEB2. '+' INDICATES INTRODUCING THE CORRESPONDING MODULE. 'PRI' MEANS PRIVATE ATTRIBUTES AND 'PUB' MEANS PUBLIC ATTRIBUTES.

Methods	On Voxceleb		On Voxceleb2	
	V-F	F-V	V-F	F-V
Baseline	89.4	86.6	80.2	81.4
+VAE	90.8	88.2	83.4	82.6
+VAE+pri	92.7	89.6	84.2	83.2
+VAE+pub	92.1	89.9	84.0	84.3
+VAE+pri+pub	**93.1**	**90.4**	**87.3**	**85.2**

TABLE V
EVALUATION ON VAE WITH EACH MODALITY IN 1:2 MATCHING TASK.

Methods	V-F	F-V
Baseline	89.4	86.6
+ Audio_VAE	89.8	87.1
+ Visual_VAE	90.3	87.7
+ (Audio + Visual)_VAE	**90.8**	**88.2**

the challenge simultaneously. We argue that more characterized private attributes and common public attributes will bring larger improvement. By jointly introducing the three components, the performance consistently improves, which verifies the effectiveness of integrating these three components into a unified framework. Our method takes advantage of private attributes to differentiate between various speakers who display inter-class similarities, which in turn results in a further increase in matching accuracy.

Evaluation on VAE. To further evaluate the effectiveness of VAE while reducing intra-class discrepancy in audio, visual data and eliminating noise in audio data, we perform ablation experiments on VAE by designing three variants: variational auto-encoder for Audio_VAE, Visual_VAE, and (Audio + Visual)_VAE, as shown in Table V. Note that all three variants are without public and private attributes. VAE contributed positively to the both modalities samples. Note that the effect of '+ Visual_VAE' is better than the effect of '+ Audio_VAE'. That is because audio clips are less affected by scene changes compared to visual images, and there is not much difference between the audio in different scenes.

Evaluation on private attributes embedding. To explore better ways of embedding private attributes, we test the impact of several different approaches in the 1:2 matching task separately, as shown in Table VI. 'Concat' means concatenating the hidden variable and private attributes simply, 'Attention mechanisms' means that the private attributes are embedded in the hidden variable using spatial attention and channel attention [47], 'Cross-transformer' means that the private attributes and the hidden variable are fused using the method shown in subsection III-C. Note that, all three variants are without public attributes guidance. All different embedding methods make a positive contribution to both V-F and F-V cases. Compared with 'Concat' and 'Attention mechanisms, 'Cross-transformer' interacts more with the private attributes and the hidden variable, obtains more discriminative features. Therefore, it achieves the best experimental results.

TABLE VI
THE RESULT OF DIFFERENT EMBEDDING METHODS IN 1:2 MATCHING TASK.

Methods	V-F	F-V
Concat	91.5	88.7
Attention mechanisms [47]	92.2	89.3
Cross-transformer [19]	**92.9**	**90.1**

E. Qualitative Results

Qualitative analysis on VAE. The attention map of the visual hidden variable extracted by the CNN backbone before VAE compared to the results after introducing VAE, as shown in Fig. 8. The ideal situation we expect is to focus on the whole face. However, due to the strong heterogeneity between modalities, the existing network learns only by VAE, and there will be cases where the network does not focus on the whole face, as shown in the first column of Fig. 8. In addition, previous works [48], [49] have shown that due to the symmetry of the face, some scenarios in which the network focuses only on one side of the face can also accomplish the identification task. Compared with using traditional CNN to extract local features, the network tends to focus on the global structure of the face and obtain more comprehensive information than bias on local information after introducing the VAE, which will help reduce the impact of intra-class discrepancy and benefit to the forthcoming audio-visual data matching.

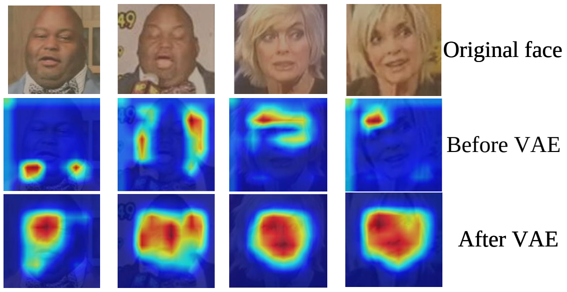


Fig. 8. Attention map of the model. The heavier the color, the more the network focuses on this area.

Qualitative analysis on public and private attributes. To better demonstrate the effectiveness of public and private attributes, we follow AML [11] to qualitatively demonstrate two matching results. As shown in Fig. 9, we have selected examples that are similar but have different public and private information. Baseline [11] and DIMNet [10] tend to mismatch the images with similar appearance while with different public (nationality) and private (nose and lips style) information. Compared to only using adversarial learning, we aggregate

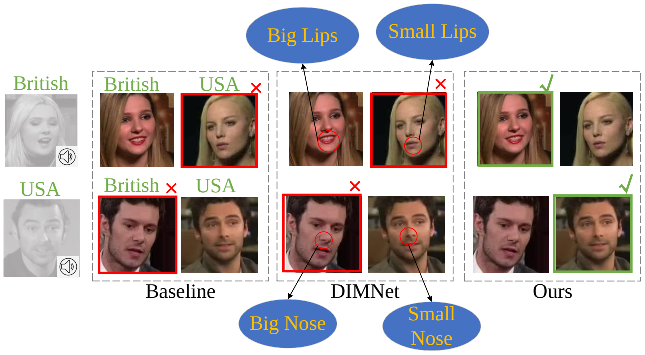


Fig. 9. Qualitative results of audio-visual cross-modal matching of the proposed method comparing to Baseline [11], DIMNet [10] in V-F challenge with $k = 2$. Red and green boxes represent wrong and correct matches respectively.

cross-modal data with the same nationality and gender by mapping them individually to their public attributes to mitigate the cross-modal heterogeneity. Compared to most existing methods which mainly employ the cross-entropy loss supervised by identity labels, private attributes reduce inter-class similarity by providing discriminative fine-grained features. Our method successfully hit the correct matching by considering both public and private attributes. This demonstrates the effectiveness of public and private attributes.

F. Parameter Analysis

There are there hyper-parameters in Eq. (18), λ_1, λ_2 λ_3 and λ_4 controlling the contribution of the VAE, identity classifier, gender attribute and nation attribute to the overall loss function, respectively. We evaluate the effect of these hyper-parameters in our method with the 1:2 matching task by separately adjusting λ_1, λ_2 λ_3 and λ_4 while fixing the other parameter. As shown in Table VII, our model is insensitive to parameters in general. It obtains the best performance when the ratio of λ_1 and λ_2 and λ_3 and λ_4 is 1.5:2:1:1.5.

TABLE VII
HYPER-PARAMETER ANALYSIS IN 1:2 MATCHING TASK.

Param	Settings	V-F	F-V	Param	Settings	V-F	F-V
λ_1	1	92.2	88.9	λ_3	0.5	92.8	90.1
	1	**93.1**	**90.4**		1	93.0	**90.4**
	1.5	91.8	88.7		1.5	**93.1**	90.3
λ_2	1	91.5	88.8	λ_4	1	**93.1**	90.4
	2	**93.1**	**90.4**		1.5	92.8	**90.5**
	3	92.6	89.3		2	92.6	89.8

V. CONCLUSIONS

To our knowledge, this paper is the first work to use variational adversarial learning-based public-private attributes for audio-visual cross-modal matching. We propose a new network called P^2**VANet** which includes a variational adversarial module, a public attributes guidance module, a private attributes embedding module, and a metric learning module to solve the problem of heterogeneous issues. VAE reduces intra-class discrepancy by constraining the hidden variable to learn inherent global information which is independent of scene changes. Public attributes mitigate cross-modal heterogeneity by providing guidance on the supervision of high-level semantic information. Private attributes reduce inter-class similarity by providing discriminative fine-grained features for distinguishing identities. Comprehensive experimental evaluation demonstrates the promising performance of the proposed P^2**VANet** while handling audio-visual cross-modal matching tasks. However, The accuracy of the attributes determines the model's performance, indicating a need to enhance attribute recognition and explore cross-modal attribute correlation. Therefore, in future research, we will design more intelligent attribute recognition networks to guide attribute embedding and explore cross-modal attribute associations through generative models (e.g., diffusion models) to enhance semantic correlation.

REFERENCES

[1] 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. 1, 3

[2] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8427–8436, 2018. 1, 2, 3, 8, 9

[3] 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. 1

[4] Y. He, X. Xu, J. Zhang, F. Shen, Y. Yang, and H. T. Shen, "Modeling two-stream correspondence for visual sound separation," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 3291–3302, 2022. 1

[5] X. Liu, E. Lakomkin, K. Vougioukas, P. Ma, H. Chen, R. Xie, M. Doulaty, N. Moritz, J. Kolar, S. Petridis, M. Pantic, and C. Fuegen, "Synthvsr: Scaling up visual speech recognition with synthetic supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18806–18815, 2023. 1

[6] G.-N. Dong, C.-M. Pun, and Z. Zhang, "Temporal relation inference network for multimodal speech emotion recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 6472–6485, 2022. 1

[7] S. Liu, S. Li, and H. Cheng, "Towards an end-to-end visual-to-raw-audio generation with gan," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 3, pp. 1299–1312, 2021. 1

[8] M. Liu, J. Wang, X. Qian, and H. Li, "Audio-visual temporal forgery detection using embedding-level fusion and multi-dimensional contrastive loss," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 1

[9] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," *arXiv preprint arXiv:1706.08612*, 2017. 1, 8, 9

[10] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," *arXiv preprint arXiv:1807.04836*, 2018. 1, 3, 4, 8, 9, 10, 11

[11] 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*, 2021. 1, 2, 3, 7, 8, 9, 10, 11

[12] 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/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16347–16356, 2021. 1, 8, 9

[13] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013. 2, 5

[14] M. Sadeghi and X. Alameda-Pineda, "Mixture of inference networks for vae-based audio-visual speech enhancement," *IEEE Transactions on Signal Processing*, vol. 69, pp. 1899–1909, 2021. 2, 4

[15] G. Chen, D. Zhang, T. Liu, and X. Du, "Self-lifting: A novel framework for unsupervised voice-face association learning," in *Proceedings of the International Conference on Multimedia Retrieval*, p. 527–535, 2022. 2

[16] 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*, 2022. 2, 3, 8, 9

[17] G. Kang, L. Zheng, Y. Yan, and Y. Yang, "Deep adversarial attention alignment for unsupervised domain adaptation: the benefit of target expectation maximization," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 401–416, 2018. 2, 7

[18] K. Cheng, X. Liu, Y. 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 28th ACM International Conference on Multimedia*, pp. 448–455, 2020. 2, 3, 7

[19] J. Zhang, Y. Xie, W. Ding, and Z. Wang, "Cross on cross attention: Deep fusion transformer for image captioning," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 8, pp. 4257–4268, 2023. 2, 4, 6, 10

[20] S. Nawaz, M. K. Janjua, I. Gallo, A. Mahmood, and A. Calefati, "Deep latent space learning for cross-modal mapping of audio and visual signals," in *Proceedings of the Digital Image Computing: Techniques and Applications*, pp. 1–7, 2019. 3

[21] 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 *2022 IEEE International Conference on Data Mining (ICDM)*, pp. 648–655, 2022. 3, 8, 9

[22] M. S. Saeed, S. Nawaz, M. H. Khan, M. Zaigham Zaheer, K. Nandakumar, M. H. Yousaf, and A. Mahmood, "Single-branch network for multimodal training," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. 3, 8, 9

[23] J. Xu, B. Liu, and Y. Xiao, "A variational inference method for few-shot learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 269–282, 2023. 4

[24] A. Makhzani, J. Shlens, N. Jaitly, I. Goodfellow, and B. Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015. 4

[25] C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel, "The variational fair autoencoder," *arXiv preprint arXiv:1511.00830*, 2015. 4

[26] Y. Burda, R. Grosse, and R. Salakhutdinov, "Importance weighted autoencoders," *arXiv preprint arXiv:1509.00519*, 2015. 4

[27] X. Chen, D. P. Kingma, T. Salimans, Y. Duan, P. Dhariwal, J. Schulman, I. Sutskever, and P. Abbeel, "Variational lossy autoencoder," in *5th International Conference on Learning Representations, ICLR (poster)*, 2017. 4

[28] A. Vahdat and J. Kautz, "Nvae: A deep hierarchical variational autoencoder," *Advances in neural information processing systems*, vol. 33, pp. 19667–19679, 2020. 4

[29] H. Li, C. Li, A. Zheng, J. Tang, and B. Luo, "Attribute and state guided structural embedding network for vehicle re-identification," *IEEE transactions on image processing*, vol. 31, pp. 5949–5962, 2022. 4, 7

[30] T. Chai, Z. Chen, A. Li, J. Chen, X. Mei, and Y. Wang, "Video person re-identification using attribute-enhanced features," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7951–7966, 2022. 4

[31] X. Cheng, M. Jia, Q. Wang, and J. Zhang, "A simple visual-textual baseline for pedestrian attribute recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6994–7004, 2022. 4

[32] S. Chen, X. Zhu, Y. Yan, S. Zhu, S.-Z. Li, and D.-H. Wang, "Identity-aware contrastive knowledge distillation for facial attribute recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 4

[33] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008. 5

[34] H. Liu, P. Sun, J. Zhang, S. Wu, Z. Yu, and X. Sun, "Similarity-aware and variational deep adversarial learning for robust facial age estimation," *IEEE Transactions on Multimedia*, vol. 22, no. 7, pp. 1808–1822, 2020. 5

[35] J. Deng, J. Guo, J. Yang, A. Lattas, and S. Zafeiriou, "Variational prototype learning for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11906–11915, 2021. 5

[36] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. 6

[37] 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. 6, 8

[38] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "Cswin transformer: A general vision transformer backbone with cross-shaped windows," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12114–12124, 2022. 6

[39] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The munich versatile and fast open-source audio feature extractor," in *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1459–1462, 2010. 6

[40] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *10th Annual Conference of the International Speech Communication Association*, pp. 312–315, 2009. 6

[41] G. Hu, Y. Hua, Y. Yuan, Z. Zhang, Z. Lu, S. S. Mukherjee, T. M. Hospedales, N. M. Robertson, and Y. Yang, "Attribute-enhanced face recognition with neural tensor fusion networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3744–3753, 2017. 7

[42] L. Zhang, L. Zuo, B. Wang, X. Li, and X. Zhen, "Variational hyperparameter inference for few-shot learning across domains," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 7448–7459, 2022. 7

[43] P. Dai, R. Ji, H. Wang, Q. Wu, and Y. Huang, "Cross-modality person reidentification with generative adversarial training," in *Proceedings of International Joint Conferences on Artificial Intelligence*, p. 2, 2018. 7

[44] H. Oh Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016. 7

[45] R. Wang, X. Liu, Y. Cheung, K. Cheng, N. Wang, and W. Fan, "Learning discriminative joint embeddings for efficient face and voice association," in *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1881–1884, 2020. 8, 9

[46] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," *arXiv preprint arXiv:1806.05622*, 2018. 8, 9

[47] Z. Ma, J. Dong, Z. Long, Y. Zhang, Y. He, H. Xue, and S. Ji, "Fine-grained fashion similarity learning by attribute-specific embedding network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 11741–11748, 2020. 10

[48] Y. Guo, H. Wang, L. Wang, Y. Lei, L. Liu, and M. Bennamoun, "3d face recognition: Two decades of progress and prospects," *ACM Computing Surveys*, vol. 56, no. 3, pp. 1–39, 2023. 10

[49] F. Yin, Y. Zhang, X. Wang, T. Wang, X. Li, Y. Gong, Y. Fan, X. Cun, Y. Shan, C. Oztireli, *et al.*, "3d gan inversion with facial symmetry prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 342–351, 2023. 10

Aihua Zheng received the B.Eng. and the integrated master's and Ph.D. degrees in computer science and technology from Anhui University, China, in 2006 and 2008, respectively, and the Ph.D. degree in computer science from the University of Greenwich, U.K., in 2012. She is currently a Professor of artificial intelligence with Anhui University. Her current research interests include computer vision and artificial intelligence, especially on person/vehicle re-identification, audio-visual learning, and multi-modal and cross-modal learning.

Fan Yuan received his B.Eng. degree in computer science and technology in 2020 from Hefei University, Hefei, China. And he received his M.S. degree in computer science and technology in 2023 from Anhui University . His current research interests include audio-visual matching and cross-modal learning.

Haichuan Zhang received his B.Eng. degree in computer science and technology in 2021 from Anhui University, Hefei, China. He is currently pursuing a Master degree in computer science and technology at Anhui University. His current research interests include multi-modal learning and self-supervised learning.

Jiaxiang Wang received his M.S. degree in control engineering in 2020 from Anhui University, Hefei, China. He is currently pursuing a Ph.D. degree in computer science and technology at Anhui University. His current research interests include pattern recognition, multi-modal learning, and computer vision.

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

Chao Tang received his M.E. degree from the School of Computer and Information Technology, Shanxi University, in 2009, and his Ph.D. degree from the School of Information Science and Technology, Xiamen University, China, in 2014. In 2016, he was a visiting scholar at Bloomfield College, U.S.A. From 2017 to 2018, he was a visiting scholar at the University of Birmingham, U.K. Since December 2018, he has been an associate professor at the School of Artificial Intelligence and Big Data, Hefei University, China. His research interests include machine learning and computer vision.