Detach and Enhance: Learning Disentangled Cross-modal Latent Representation for Efficient Face-Voice Association and Matching

Zhenning Yu[1,2], Xin Liu[1,2,3,*], Yiu-ming Cheung[3,*], Minghang Zhu[1,4], Xing Xu[5], Nannan Wang[6], Taihao Li[2]
[1]Dept. of Comput. Sci. & Fujian Key Lab. of Big Data Intelligence and Security, Huaqiao University, Xiamen, China
[2]Zhejiang Lab, Hangzhou, China
[3]Dept. of Comput. Sci. and Institute of Research and Continuing Education, HK Baptist University, Hong Kong SAR, China
[4]Xiamen Key Lab. of Computer Vision and Pattern Recognition, Huaqiao University, Xiamen, China
[5]Dept. of Computer Sci. and Eng., University of Electronic Science and Technology of China, Chengdu, China.
[6]State Key Lab. of Integrated Services Networks & School of Telecommun. Eng., Xidian University, Xi'an, China
{zny, xliu, mhz}@hqu.edu.cn, ymc@comp.hkbu.edu.hk, xing.xu@uestc.edu.cn, nnwang@xidian.edu.cn, lith@zhejianglab.com

2022 IEEE International Conference on Data Mining (ICDM) | 978-1-6654-5099-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICDM54844.2022.00075

***Abstract*—Many researches in cognitive science have shown that humans often perform face-voice association for various perception tasks, and some recent data mining works have been designed in emulating such ability intelligently. Nevertheless, most methods often suffer from the degraded performance when there exist semantically irrelevant interference factors across different modalities. To alleviate this concern, this paper presents an efficient Disentangled Cross-modal Latent Representation (DCLR) method to adaptively detach the discriminative feature attributes and enhance the face-voice association. To be specific, the proposed DCLR framework consists of two-stage cross-modal disentangling process. First, the former stage employs the supervised contrastive learning to push the representations of face-voice data from the same person closer while pulling those representations of different person away. Then, the latter stage freezes all the parameters of the former stage, and further innovates a multi-layer orthogonal decoupling scheme to learn the disentangled latent representations, while filtering out the modality-dependent irrelevant factors. Besides, the cross-modal reconstruction loss is further utilized to narrow down the semantic gap between heterogeneous feature expressions. Through the joint exploitation of the above, the proposed framework can well associate the face-voice data to benefit various kinds of cross-modal perception tasks. Extensive experiments verify the superiorities of the proposed face-voice association framework and show its competitive performances.**

***Index Terms*—Face-voice association, disentangled latent representation, contrastive learning, orthogonal decoupling**

I. Introduction

Many cognitive researches lend credence to the hypothesis that humans are able to hear voices of known individuals to form mental pictures of their facial appearances, and vice versa [1]. Even, some neurologic studies have shown that humans also are able to associate the voices of unknown individuals to the relevant pictures of their faces. In recent years, face and voice have proven to be the most valuable media data for representing the biometric identity information, which can greatly help recognize, search and organize human identities in artificial intelligence system.

Face-voice association is a task of finding their semantic correspondence, which is of crucial importance to creating natural human machine interaction systems and benefiting many real applications, such as active speaker annotation and diarization [2]. For instance, when watching an unvoiced TV show, an intelligent media annotation system embedded with face-voice association technique can well assign an appropriate voice actor to a matched performer. In recent years, much effort has been paid to design data mining and machine intelligence algorithm for recognizing voice-face associations [3], [4], and some researches have been developed. Intuitively, a natural way is to learn a common feature embedding to minimize the heterogeneity between the face images and voice segments, whereby the mapping features in such space can be directly measured for face-voice association. For instance, LAFV [5] leverage the triplet network to learn the co-embedding of modality-representations of human faces and voices, while CME [6] train a two-branch neural network to learn cross-modal embeddings between face images and voices.

Despite these works are able to correlate the relevant face-voice data, their performances are far from the expectation. To the best of our knowledge, the study of efficient face-voice association is still under early stage and there still face three challengings: 1) **Modality heterogeneity**: face and audio samples are captured by different sensors, and there exist huge modality differences between face and voice features. 2) **Weak semantic correlation**: most existing face-voice association works often extract the modality-specific feature vectors from single modality data, which inherently ignores the latent shared

This work was supported in part by Open Research Projects of Zhejiang Lab under Grants 2021KH0AB01 and 2021KG0AB01, in part by National Natural Science Foundation of China under Grants 61976049, 61922066 and 61876142, in part by Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15, in part by National Science Foundation of Fujian Province under Grant 2020J01084, in part by NSFC/RGC Joint Research Scheme under Grant N_HKBU214/21, General Research Fund of RGC under Grant 12201321, and Hong Kong Baptist University under Grant: RC-FNRA-IG/18-19/SCI/03. Corresponding authors: Xin Liu and Yiu-ming Cheung.

978-1-6654-5099-7/22/$31.00 ©2022 IEEE
DOI 10.1109/ICDM54844.2022.00075

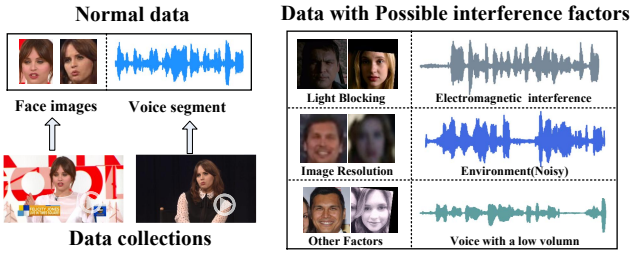


Fig. 1. Normal data and examples with interference factors.

information and their association and correlation performances need further improvements. 3) **Unexpected interference factors**: there exist different kinds of interference factors during the data collections, which often bring significant difficulties to semantically correlation the heterogeneous face and voice data. As shown in Fig. 1, the face image is often influenced by different lighting condition, while voice data may be corrupted by the environmental noise and electromagnetic interference. Note that, these complex factors often make it difficult to learn the reliable face-voice associations.

In this paper, we hypothesize that the latent cross-modal representation among the face and voice should be robust against to the interference factors. Towards this end, we learn an efficient Disentangled Cross-modal Latent Representation (DCLR) to disentangle the common embeddings from the face-voice data, which consists of two-stage learning process. The former stage employs the supervised contrastive learning to enlarge the distance margin between the positive and negative face-voice pairs, while the latter stage aims to learn the disentangled latent representations that are shared across the face and voice data. The main contributions are four-fold:

- A novel disentangled cross-modal latent representation framework is explicitly designed to learn the face-voice association, which ensures that the learnt cross-modal embeddings are more effective for various matching tasks.
- The designed cross-modal batch is able to significantly improve the data utilization for face-voice association.
- An efficient multi-layer orthogonal decoupling scheme is addressed to learn the common embdeddings, while filtering out the interference factors within each modality.
- Extensive experiments verify the advantages of DCLR under various face-voice matching scenarios.

The rest of this paper is organized as follows: Section II briefly surveys the face-voice association works, and Section III elaborates the proposed learning framework in detail. The extensive experiments and comparisons are introduced in Section IV. Finally, we draw a conclusion in Section V.

II. RELATED WORK

The key issue of face-voice association is to learn their semantic correlations, and various data mining works have been developed. An intuitive way is to learn a common embeddings for correlating the voices and faces so that they can be compared with each other. The pioneer SVHF [7] method utilizes CNN architectures to learn the joint representation of voices and faces, and formulates their association problem as a binary classification problem. It is found that this framework is not flexible enough for different cross-modal face-voice matching tasks. Benefiting from the advances in multi-modal deep learning, they further form the positive and negative face-voice pairs acquired from the same talking face in a video, and attempt the contrastive loss to minimize the distance between the embeddings of positive or negative pairs [8]. In addition, LAFV [5] utilizes the triple network to learn the co-embedding of human faces and voices, while CME [6] exploits the N-pair loss to learn the cross-modal embeddings between face and voice data. With more semantic information, DIMNet [3] employs multi-label classification to learn a shared representation by mapping face and voice individually to their common covariates, *i.e.*, identity, nationality and gender. Similarly, SSNet [9] learns a shared deep latent space representation of multi-modal information, and leverage the class centers to eliminate the pairwise or triplet supervision for face-voice association. Until recently, Wang's [4] attempts to learn discriminative joint embedding by using bi-directional ranking constraint, identity constraint and center constraint for face-voice association. Note that, these two models just utilize the identity constraint to supervise the common embedding learning, which cannot fully exploit the shared latent semantic information for reliable cross-modal correspondence.

Disentangled representation learning aims to design the appropriate objective functions to learn the disentangled representations from the raw data [10]. This not only improves the interpretability of the model, but also enables analysis for specific elements. Variational auto-encoder (VAE) [11] is designed to solve factor analysis on simple datasets. Beta-vae (β-VAE) [12] upgrades this process by introducing a hyperparameter β that represents the degree of decoupling. Recently, disentangled representation learning has also been applied in the field of rare event detection [13]. In the filed of cross-modal face-voice association learning, CMBM [14] considers the disentangled identity factors as the pure identity information to bridge the face and voice data, and construct a disentangled learning module to align the identity information between the paired face and voice data. Experimentally, this approach has shown its outstanding performance on face-voice matching task. Nevertheless, it is found that this approach does not fully consider the distance margin between the positive face-voice pair and negative face-voice pair, thereby the derived common cross-modal embeddings are not discriminative enough for better association performance.

III. METHODOLOGY

The overall concept of the proposed face-voice association framework is illustrated in Fig. 2, and this section elaborates the proposed two-stage learning process in detail.

A. Problem Description

Without loss of generality, let $\mathbf{X}_f=\{\mathbf{f}_k\}_{k=1}^N$ and $\mathbf{X}_v=\{\mathbf{v}_k\}_{k=1}^N$, respectively, represent the face and voice

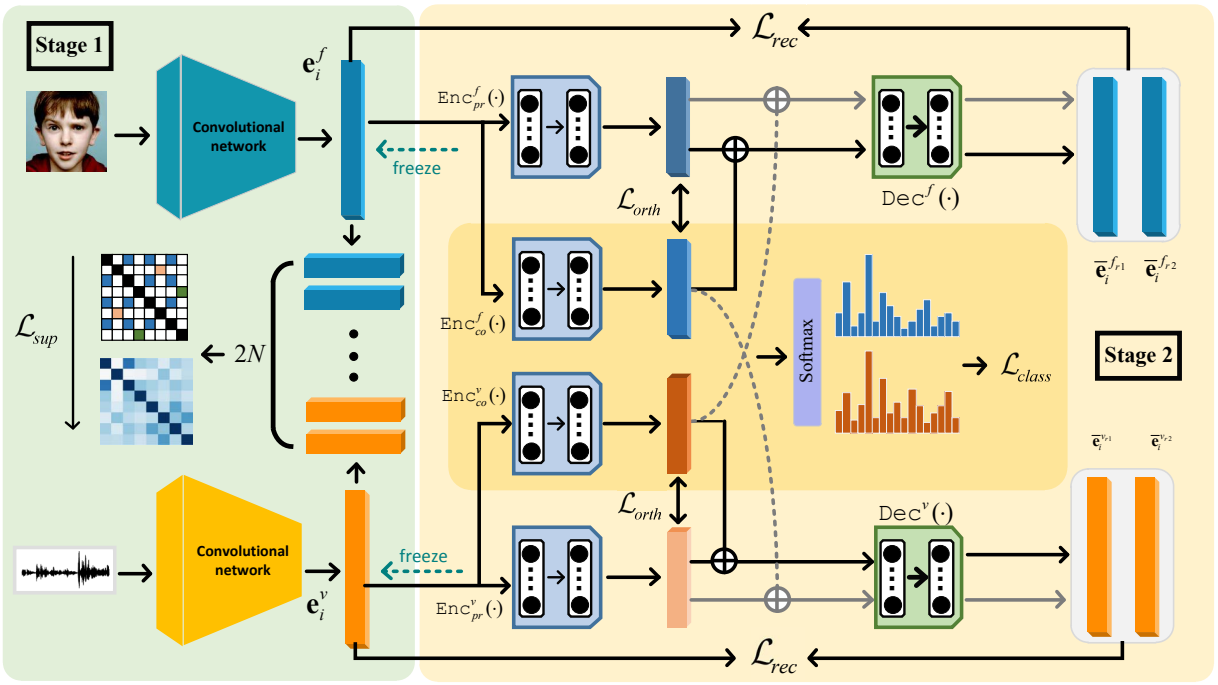


Fig. 2. The schematic architecture of the proposed face-voice association learning framework.

data, with $\mathbf{f}_i$ and $\mathbf{v}_i$ being the i-th face-voice data pair, where N is a number of the training dataset size. The feature extraction backbone is shown in Fig. 2, which is a dual-brach network consisting of face subnetwork and voice subnetwork, respectively, denoted as $\mathbf{e}_i^f = \texttt{fnet}(\mathbf{f}_i) \in \mathbb{R}^{d_1}$ and $\mathbf{e}_i^v = \texttt{vnet}(\mathbf{v}_i) \in \mathbb{R}^{d_1}$, where d_1 denote the initial dimensionality. To disentangle the shared latent common embedding, we construct four encoders, termed as face common encoder $\texttt{Enc}_{co}^f(\cdot)$, face private encoder $\texttt{Enc}_{pr}^f(\cdot)$, voice common encoder $\texttt{Enc}_{co}^v(\cdot)$, voice private encoder $\texttt{Enc}_{pr}^v(\cdot)$, and two decoders $\texttt{Dec}^f(\cdot)$ and $\texttt{Dec}^v(\cdot)$. Specifically, the common encoders are utilized to learn the disentangled feature vectors that are shared across the face and voice data, while the private encoders are employed to filter out the modality-dependent factors with respect to each face and voice data. The decoders are employed to bridge the semantic gap between the inputs and outputs.

B. Supervised Contrastive Learning

Contrastive learning can be utilized for better cross-modal representation learning, which allows the model to contrast the positive pairs from sets of negative samples [15], [16]. To associate the face and voice data, the semantically similar face-voice pairs should have shared representations, and vice versa. To this end, we construct a cross-modal batch instead of directly using a mini-batch, and the face-voice data is similar if they are collected from the same person and close to each other in the shared representation space. Given a set of face-voice pairs $\mathbf{z}_k = \{\mathbf{e}_k^f, \mathbf{e}_k^v\}_{k=1\cdots n}$, and their identity label $\{\mathbf{y}_k\}_{k=1\cdots n}$, we construct a cross-modal batch $\{\mathbf{z}_i, \bar{\mathbf{y}}_i\}_{i=1\cdots 2n}$, with mapping relationship by $\mathbf{z}_k = \mathbf{e}_k^f, \mathbf{z}_{k+n} = \mathbf{e}_k^v$ and the label $\bar{\mathbf{y}}_k = \bar{\mathbf{y}}_{k+n} = \mathbf{y}_k, k = 1\cdots n$. By this combination, we normalize $\mathbf{z}_k$ to compute the similarity between different data items. Suppose the index i is the anchor sample, the supervised contrastive loss is defined as:

$$\mathcal{L}_{sup} = \sum_{i=1}^{2n} \frac{-1}{|P(i)|} \sum_{p \in P(i)} \log \frac{\exp(\mathbf{z}_i \cdot \mathbf{z}_p / \tau)}{\sum_{a \in A(i)} \exp(\mathbf{z}_i \cdot \mathbf{z}_a / \tau)}, \quad (1)$$

where $P(i) = \{p \in A(i) : \bar{\mathbf{y}}_p = \bar{\mathbf{y}}_i\}$ represents the set of indices of all positive samples with the i-th identity, $|P(i)|$ denotes the count of the positive examples, $A(i)$ denote the data set that exclude itself and τ is a scalar temperature parameter.

C. Disentangled Cross-modal Latent Representation

The supervised contrastive learning stage is able to push the representations of face-voice data from the same person closer while pulling those representations of different person away. Nevertheless, the derived feature representations not only contain the identity discrimination, but also include the interference factors that may bring the negative impact to the correlation mining. To alleviate such concern, we further freeze all parameters of the first stage, and propose a a disentangled cross-modal representation learning method to promote the face-voice association learning. To be specific, we explore the disentangled latent representations among the face and voice data, and simultaneously attempt to filter out the interference factors within each modality. Within this module, identity association learning, orthogonalized disentangling strategy and cross-modal reconstruction scheme are

seamlessly embedded to promote the disentangled cross-modal latent representation learning process.

Identity Association Learning Module: It is noted that the semantic constraints are usually added to ensure the common representations from the same person to be as similar as possible. Given the k-th face-voice data pair $\{\mathbf{e}_k^f, \mathbf{e}_k^v\}$, we extract the face common embedding and voice common embedding, respectively, by $\mathbf{e}_k^{f_c}{=}\mathtt{Enc}_{co}^f(\mathbf{e}_k^f){\in}\mathbb{R}^{d_2}$, and $\mathbf{e}_k^{v_c}{=}\mathtt{Enc}_{co}^v(\mathbf{e}_k^v){\in}\mathbb{R}^{d_2}$, where d_2 is the output dimensionality of the common encoder network. For the embedding data pair $\{\mathbf{e}_k^{f_c}, \mathbf{e}_k^{v_c}, \mathbf{y}_k\}_{k=1\cdots n}$, we feed them into to a single fully connected layer, and utilize the cross-entropy loss to optimize the identity classification result:

$$\mathcal{L}_{class} = \frac{1}{n}\sum_{k=1}^{n} H(\mathbf{e}_k^{f_c}, \mathbf{y}_k) + H(\mathbf{e}_k^{v_c}, \mathbf{y}_k), \tag{2}$$

$$H(\mathbf{e}_k, \mathbf{y}_k) = -\sum_{c=1}^{M} (\mathbf{y}_k{=}c)\log(\phi(\mathbf{x}_k)_c), \tag{3}$$

where M represents the number of all identities, $\phi(\cdot)_c$ indicates a single fully connected layer to obtain the probability of x belonging to the semantic label c. The identity label $y_k(k = 1...n)$ is same for each input embedding pairs.

Multi-layer Disentangling Module: To filter out the irrelevant factors that existed in each modality, we further extract the private embedding by $\mathbf{e}_k^{f_p}{=}\mathtt{ReLU}(\mathtt{Enc}_{co}^f(\mathbf{e}_k^f))$ and $\mathbf{e}_k^{v_p}{=}\mathtt{ReLU}(\mathtt{Enc}_{pr}^f(\mathbf{e}_k^v))$ to separate the interference factors, where $\mathtt{ReLU}$ is utilized to regularize the output of the embedding network. For disentangled representation learning [10], its main objective is to completely separate the common embedding vector and private embedding vector from the input feature representation. To this end, we employ multi-layer orthogonality constraints to increase their heterogeneity, and the corresponding orthogonal loss function is defined as:

1) **Instance-level disentangling operation**: For each modality data, the common embedding and private embedding should be disentangled to be irrelevant to each other, and the cosine similarity (CS) can be utilized to measure the magnitude of the difference. Often, a smaller cosine similarity represents a lower correlation between the common embedding and private embedding. Accordingly, the the orthogonal loss based on CS is defined as:

$$\mathcal{L}_{orth}^{CS} = \frac{1}{n}\sum_{k=1}^{n}[\cos(\mathbf{e}_k^{f_c}, \mathbf{e}_k^{f_p}) + \cos(\mathbf{e}_k^{v_c}, \mathbf{e}_k^{v_p})], \tag{4}$$

where $\cos(.,.)$ represents the cosine similarity between two input embeddings. By minimizing such loss, the model encourages orthogonality between the common embedding and private embedding, which can be well utilized to disentangle the shared common embedding across face and voice data. For the convenience of explanation, we take the face modality as an example and the similar observation goes for the voice modality. Suppose two embeddings e^{f_c} and e^{f_p} are derived respectively from the face common and private encoders, and we normalize each embeddings to lie on the unit hypersphere by ℓ_2 normalization:

$$e^{f_c} = (\frac{\mathbf{c}_1^f}{||\mathbf{e}_k^{f_c}||_2} + \frac{\mathbf{c}_2^f}{||\mathbf{e}_k^{f_c}||_2} + ... + \frac{\mathbf{c}_{d_2}^f}{||\mathbf{e}_k^{f_c}||_2}), \tag{5}$$

$$e^{f_p} = (\frac{\mathbf{p}_1^f}{||\mathbf{e}_k^{f_p}||_2} + \frac{\mathbf{p}_2^f}{||\mathbf{e}_k^{f_p}||_2} + ... + \frac{\mathbf{p}_{d_2}^f}{||\mathbf{e}_k^{f_p}||_2}), \tag{6}$$

where $\mathbf{c}_i^f$ and $\mathbf{p}_i^f$, respectively, represent the i-th element in common embedding and private embedding, d_2 represents the dimension of the feature embedding. Accordingly, the orthogonal loss function can be expressed as:

$$\begin{aligned}\cos(\mathbf{e}^{f_c}, \mathbf{e}^{f_p}) &= \frac{\mathbf{e}^{f_c}\cdot\mathbf{e}^{f_p}}{||\mathbf{e}^{f_c}||\times||\mathbf{e}^{f_p}||} \\ &= (\mathbf{c}_1^f\mathbf{p}_1^f + \mathbf{c}_2^f\mathbf{p}_2^f + ... + \mathbf{c}_{d_2}^f\mathbf{p}_{d_2}^f).\end{aligned} \tag{7}$$

Within the derived latent space, it is reasonable to assume that each element in the feature vector represents a different semantic factor corresponding to the identity information. In order to maximally preserve the identity information, the disentangling process can guarantee that the linear independence between the common embeddings and private embeddings is measured by means of inner product.

2) **Modality-layer disentangling operation**: For each modality instance, the instance-level disentangling operation can make the common embedding and private embedding of each instance to be irrelevant to each other. For the training process, the data instances often interacts with each other, and it is imperative to filter out the interference factors within each modality and simultaneously consider the semantic identity information to distinguish each modality. Accordingly, we explicitly separate the common embedding space and the private embedding space during the learning process. Let $F_c{=}\{\mathbf{e}_k^{f_c}\}_{k=1}^n$, $F_p{=}\{\mathbf{e}_k^{f_p}\}_{k=1}^n$, $V_c{=}\{\mathbf{e}_k^{v_c}\}_{k=1}^n$ and $V_p{=}\{\mathbf{e}_k^{v_p}\}_{k=1}^n$, respectively, denote the common feature matrix and private feature matrix, whose rows are respectively the common and private representations for the instances of face and voice modality. Then, we employ modality constraint (MC) to encourage orthogonality between the representations in the common and private feature embeddings:

$$\mathcal{L}_{orth}^{MC} = ||F_c{F_p}^{\mathrm{T}}||_F + ||V_c{V_p}^{\mathrm{T}}||_F, \tag{8}$$

where $||\cdot||_F$ represent Frobenius norm. This loss function measures the degree of association between common and private embeddings in each modality.

3) **Mutual-level disentangling operation**: Instance-level disentangling operation and modality-layer disentangling operation encourage different ways of orthogonalization within each data instance and each modalities. To aggregate these disentangling capability, we propose a new orthogonal aggregated method by using $\mathtt{AND}$ gate and $\mathtt{OR}$ gate. Different from the above two operations that directly calculate feature embedding, here it explores overall instances of cross-modal data in the cross-modal latent space. To be specific, we employ

AND gate to group common embeddings and OR gates to combine private embeddings:

$$e_k^A = \texttt{AND}(\mathbf{e}_k^{f_c}, \mathbf{e}_k^{v_c}), \quad e_k^O = \texttt{OR}(\mathbf{e}_k^{f_p}, \mathbf{e}_k^{v_p}), \tag{9}$$

where AND represents the gate of AND operation, OR denotes the gate of OR operation. Through the disentangling aggregation, we can obtain the discriminative common embeddings and private embeddings. The AND gate and OR gate are implemented by multiplication and addition, respectively. Specifically, let $R^A{=}\{e_k^A\}_{k=1}^n{\in}\mathbb{R}^{n\times d_2}$ and $R^O{=}\{e_k^O\}_{k=1}^n{\in}\mathbb{R}^{n\times d_2}$, respectively, denote the embedding matrices whose rows are the common and private embeddings derived from the cross-modal gate operations, then the orthogonal loss based on AO is defined as:

$$\mathcal{L}_{orth}^{AO} = ||R^A(R^O)^{\mathrm{T}}||_F, \tag{10}$$

Through the minimizing of Eq. (10), the disentangled cross-modal latent space supports orthogonality across face and voice data, and the gates operation also contributes to the elimination of cross-modal heterogeneity.

4) **Integrated disentangling loss**: For efficient face-voice association, the regularization of different orthogonal layers should be exploited in an integrated way, and the following objective function is utilized to learn the fine-grained face-voice association:

$$\mathcal{L}_{orth}^{all} = \mathcal{L}_{orth}^{CS} + \mathcal{L}_{orth}^{MC} + \mathcal{L}_{orth}^{AO}. \tag{11}$$

Cross-modal Reconstruction Module: The combination of common embedding and private embedding enhances the robustness of the reconstruction results. Meanwhile, the refined embedding between different layers should semantically match the input embedding to maintain the semantic consistency. To achieve cross-modal association, we perform modality-specific reconstruction and cross-modal reconstruction in tandem. For the modality-specific reconstruction, the decoders of common embeddings and the private embeddings within each modality are combined to reconstruct the input representations as:

$$\bar{\mathbf{e}}_k^{f_{r_1}} = \texttt{Dec}^f(\mathbf{e}_k^{f_c} \oplus \mathbf{e}_k^{f_p}), \quad \bar{\mathbf{e}}_k^{v_{r_1}} = \texttt{Dec}^v(\mathbf{e}_k^{v_c} \oplus \mathbf{e}_k^{v_p}), \tag{12}$$

where $\oplus$ represents the addition of two embedding vectors, $\bar{\mathbf{e}}_k^{f_{r_1}}{\in}\mathbb{R}^{d_1}$ and $\bar{\mathbf{e}}_k^{v_{r_1}}{\in}\mathbb{R}^{d_1}$ are of the same dimensionality. For the cross-modal reconstruction, we mutually reconstruct the face embedding and voice embedding in a cross-modal way to maintain the semantic consistency:

$$\bar{\mathbf{e}}_k^{f_{r_2}} = \texttt{Dec}^f(\mathbf{e}_k^{v_c} \oplus \mathbf{e}_k^{f_p}), \quad \bar{\mathbf{e}}_k^{v_{r_2}} = \texttt{Dec}^v(\mathbf{e}_k^{f_c} \oplus \mathbf{e}_k^{v_p}), \tag{13}$$

Consequently, we integrate the reconstruction loss to ensure the semantic consistency between face and voice:

$$\begin{aligned} \mathcal{L}_{rec} \quad &= \tfrac{1}{n}\sum_{k=1}^{n}[f_{rec}(\mathbf{e}_k^f, \bar{\mathbf{e}}_k^{f_{r_1}}) + f_{rec}(\mathbf{e}_k^f, \bar{\mathbf{e}}_k^{f_{r_2}})] \\ &+ \tfrac{1}{n}\sum_{k=1}^{n}[f_{rec}(\mathbf{e}_k^v, \bar{\mathbf{e}}_k^{v_{r_1}}) + f_{rec}(\mathbf{e}_k^v, \bar{\mathbf{e}}_k^{v_{r_2}})], \end{aligned} \tag{14}$$

Algorithm 1 Learning algorithm of DCLR framework

Input: Training data $\mathbf{X}_f{=}\{\mathbf{f}_k\}_{k=1}^N$ and $\mathbf{X}_v{=}\{\mathbf{v}_k\}_{k=1}^N$, Temperature parameter τ, Dimension d_1, d_2, Epoch_number L.
Output:Model parameters Θ.
Initialization: Initialize face subnetwork $\texttt{fnet}(\cdot)$ and voice subnetwork $\texttt{vnet}(\cdot)$ in stage one, Initialize encoders and decoders $\texttt{Enc}_{co}^f(\cdot)$, $\texttt{Enc}_{co}^v(\cdot)$, $\texttt{Enc}_{pr}^f(\cdot)$, $\texttt{Enc}_{pr}^v(\cdot)$, $\texttt{Dec}^f(\cdot)$, $\texttt{Dec}^v(\cdot)$ in stage two.

// Stage One
for l = 1,2,...,L **do**
  Pre-train face and voice subnetworks $\texttt{fnet}(\cdot)$, $\texttt{vnet}(\cdot)$ with inputs $\mathbf{X}_f{=}\{\mathbf{f}_k\}_{k=1}^N$, $\mathbf{X}_v{=}\{\mathbf{v}_k\}_{k=1}^N$.
  Calculate the supervised contrastive loss $\mathcal{L}_{sup}$ with Eq. (1) in a cross-modal batch.
end for
// Stage Two
Freeze the subnetworks parameters
Decoupling of features inherited from stage one
for l = 1,2,...,L **do**
  Update $\texttt{Enc}_{co}^f(\cdot)$, $\texttt{Enc}_{co}^v(\cdot)$ with Eq. (2) and Eq. (11) and Eq. (14)
  Update $\texttt{Enc}_{pr}^f(\cdot)$, $\texttt{Enc}_{pr}^v(\cdot)$ with Eq. (11) and Eq. (13)
  Update $\texttt{Dec}^f(\cdot)$, $\texttt{Dec}^v(\cdot)$ with Eq. (14)
  Adjusting parameters in the encoder and decoder with back propagation
end for
return The face subnetwork $\texttt{fnet}(\cdot)$, voice subnetwork $\texttt{vnet}(\cdot)$ and cross-modal latent encoder $\texttt{Enc}_{co}(\cdot)$.

where $f_{rec}(\mathbf{e}, \bar{\mathbf{e}}){=}||\mathbf{e}{-}\bar{\mathbf{e}}||_2^2$ represents mean-square error function, and the final loss function is formulated as:.

$$\mathcal{L}_{disent} = \mathcal{L}_{class} + \mathcal{L}_{orth}^{all} + \mathcal{L}_{rec}. \tag{15}$$

D. Training Strategy

The complete training process of the proposed DCLR method consists of two-stage learning procedures. The first stage constructs a cross-modal batch and optimizes the feature representation with supervised contrastive loss in Eq. (1). The second stage freezes all the parameters of the first stage, and further learns the disentangled latent representations by using loss function in Eq. (15). The whole optimization problem can be efficiently solved by Adam [17], with the batch size setting at 64, while the learning rate is fixed to be 10^{-5}. The proposed DCLR algorithm is summarized in Algorithm 1.

IV. EXPERIMENTS

As suggested in most baseline [3], [4], [7], the public available Voxceleb [18] and VGGFace dataset [19] are selected for evaluation. Voxceleb dataset consists of short videos collected from 1251 celebrity interviews, while VGGFace dataset contains 2622 identity information data. For face-voice association, we select 1225 identities with overlapping peoples within these two datasets, and exactly follow the work [7] to split data into train/vaildation/test sets without

identity overlapping. In the experiments, we make use of both still images from VGGFace and frames extracted from the videos in the VoxCeleb dataset during training. The statistical information of the data information is shown in Table I.

TABLE I
STATISTICAL NUMBERS FOR SPLITTING DATASETS.

	Train	Val	Test	total
Face images	106584	12260	20076	138920
Speech segments	106584	14182	21850	142616
Identities	924	112	189	1225

A. Implementation Details

Face extractor: The cropped RGB face image is scaled into the size of 224×224×3, and followed by preprocessing with the similar operations stated in work [8]. Meanwhile, the face sub-network backbone is implemented by the standard Inception-ResNet-v1 [20] architecture.

Voice extractor: The voice data are detected from the original video by voice activity detector, and obtained by eliminating the silent period. Accordingly, 64-dimensional log melspectrograms are generated (window size: 25ms, hop size: 10ms) and followed by mean and variance normalization. The voice sub-network is implemented by DIMNet-voice [3].

During the training process, τ is experimentally fixed to be 0.07 during the supervised contrastive learning in stage one. All of the encoders and decoders in stage two are composed of several fully connected layers, and the dimensions of these layer encoders are arranged as 1024→512→64, and each layer is followed by the ReLU activation function. Conversely, the decoder is symmetrical to the encoder, and the dimensions of these layer decoders are set at 64→512→1024. Before reconstruction, we normalize the embeddings of the two inputs, and choose the Tanh activation function instead of ReLU in the decoders to provide a diverse reconstruction effect. The evaluation protocols of face-voice association are four-fold:

1) **Verification**: Given a face and a voice data example, the purpose of cross-modal verification task is to determine whether they are collected from the the same person or not. The verification performance is evaluated with AUC quantitative indicator (area under the ROC curve).

2) **1:2 matching**: Given an anchor example from one modality and two candidates from the other modality, the objective of 1:2 matching task is to find out which candidate has the same identity information as the anchor. The performance is evaluated with the metric of accuracy (ACC) [21].

3) **1:N matching**: This task is an extension of the aforementioned 1:2 matching task, and the only difference is that there are N candidates instead of two candidates. Similarly, the accuracy is utilized to measure the matching performance.

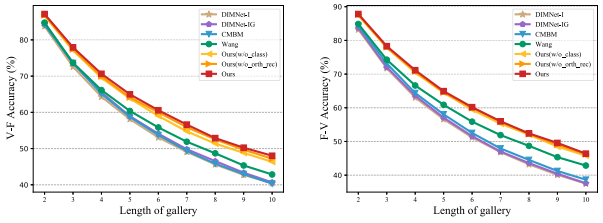
4) **Retrieval**: This task extends the matching work to cross-modal retrieval scenario. That is, given a query data of one modality, it allows to index more candidates with the same identity as matched in another modality, and its performance is evaluated with standard mean average precision (mAP) [22].

To evaluate the face-voice association performance, we divide all cross-modal association tasks into two cases: face matching voice (F-V) and voice matching face (V-F). Similar to most baseline [3], [4], [7], the symbol (G) represents that all data in the task have the same gender and the symbol (U) represents unstratified group. In the experiments, the competing SVHF [7], DIMNet [3], CMBM [14] and Wang's model [4] are selected for meaningful comparisons.

B. Performance Analysis and Comparison

Results of face-voice association: The 1:2 matching, verification, and cross-modal retrieval results about face-voice associations are shown in Table II, in which the best results of our method are marked in red, and the best results among the competing methods are marked in green. It can be clearly observed that the proposed DCLR method has always achieved the better performance over the baselines. For the 1:2 matching task and verification task, no matter the dataset is gender-restricted or unrestricted, the proposed DCLR method improves the performance by 2%-4% over the results by the competing Wang [4]. That is, the cross-modal latent representation derived from the proposed DCLR framework can provide valuable identity information to find the associations between faces and voices. For the more challenging retrieval task, it can be found that the proposed DCLR approach always yields the highest mAP values, in both V-F and F-V retrieval tasks. For instance, the mAP score obtained by the proposed DCLR approach reaches up to 6.96 on V-F task, which is significantly higher than the result 4.48 generated by CMBM [14] and 5.13 obtained by Wang [4]. That is, the proposed DCLR model is able to well correlate the semantically similar face and voices.

Besides, the representative 1:N matching results are shown in the Fig. 3, it can be also found that the proposed disentangling network almost always delivers the best results on different N values. That is, the derived disentangled latent representation is beneficial to find the potential associations between the faces and voices. Therefore, the proposed disentangling network model not only can preserve much information about the identify information across face and voice data, but also is able to filter out the irrelevant information that can promote the face-voice association performance.



(a) V→F 1:N matching task. (b) F→V 1:N matching task.

Fig. 3. Comparison of 1:N matching performance.

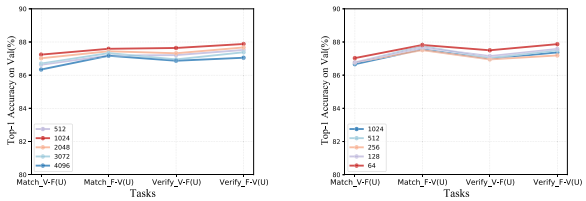
Results of different disentangling operations: Within the proposed framework, we employ the multi-layer orthogonality

TABLE II
COMPARISON WITH BASELINES ON 1:2 MATCHING TASK, VERIFICATION TASK AND RETRIEVAL TASK.

Tasks	1:2 Matching(ACC)				Verification(AUC)				Retrieval(mAP)	
Methods	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F	F-V
SVHF [7]	81.0	79.5	63.9	63.4	-	-	-	-	-	-
DIMNet-I [3]	83.77	83.20	72.88	70.94	83.21	83.44	71.93	70.61	5.10	3.89
DIMNet-IG [3]	84.39	83.70	73.15	71.67	83.10	83.56	72.07	71.14	4.68	4.01
CMBM [14]	84.17	84.35	73.64	72.21	83.83	83.89	73.18	71.71	4.48	4.13
Wang's [4]	84.76	84.87	74.10	74.22	84.25	84.87	74.63	74.74	5.13	4.45
Ours(w/o $\mathcal{L}_{sup}$)	84.03	83.61	72.06	70.09	83.85	83.92	69.88	68.76	4.39	3.87
Ours(w/o $\mathcal{L}_{class}$)	86.30	87.04	76.15	75.96	85.71	86.34	74.69	75.20	6.65	5.43
Ours(w/o $\mathcal{L}_{orth}^{all}$ + $\mathcal{L}_{rec}$)	86.68	87.35	76.85	76.99	**87.15**	**87.52**	77.22	77.08	6.80	5.84
Ours	**86.79**	**87.45**	**77.40**	**77.58**	86.76	86.89	**77.62**	**77.44**	**6.96**	**5.90**

TABLE III
COMPARISON WITH DIFFERENT ORTHOGONAL LOSS FUNCTIONS ON 1:2 MATCHING TASK, VERIFICATION TASK AND RETRIEVAL TASK.

Tasks	1:2 Matching(ACC)				Verification(AUC)				Retrieval(mAP)	
Different $\mathcal{L}_{orth}$	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F(U)	F-V(U)	V-F(G)	F-V(G)	V-F	F-V
$\mathcal{L}_{orth}^{CS}$	86.71	87.37	77.08	77.12	86.68	86.85	77.50	77.23	6.94	5.74
$\mathcal{L}_{orth}^{MC}$	86.53	87.27	77.27	77.28	86.96	87.54	75.87	75.92	6.57	5.84
$\mathcal{L}_{orth}^{AO}$	86.58	87.40	76.89	76.96	**87.18**	**87.59**	77.22	77.25	6.95	5.89
$\mathcal{L}_{orth}^{all}$	**86.79**	**87.45**	**77.40**	**77.58**	86.76	86.89	**77.62**	**77.44**	**6.96**	**5.90**



(a) Different d_1 values. (b) Different d_2 values.

Fig. 4. Impacts of d_1 and d_2 within the DCLR framework.

constraints between the common and private encoders in each modality to increase their heterogeneity. To evaluate the effectiveness of each disentangling operation, we record the results of different disentangling operations and their combinations. Table III displays the performance of the DCLR framework with different orthogonal loss functions (i.e., $\mathcal{L}_{orth}^{CS}$, $\mathcal{L}_{orth}^{MC}$, $\mathcal{L}_{orth}^{AO}$ and $\mathcal{L}_{orth}^{all}$) on the task of cross-modal face-voice association, in which the best results are marked in red. It can be observed that the $\mathcal{L}_{orth}^{all}$ embedded within the DCLR framework achieves a bright performance. In the orthogonal subspace, the combination of different orthogonal strategies provides the greatest degree of separation of interference factors. The multi-layer disentangling module scheme allows each method to complement each other's advantages in cross-modal face and voice association learning. It is worth mentioning that the gating strategy achieves the best results on verification(U) tasks. The gate operation focuses on the heterogeneity of cross-modal data and provides a useful complement to the verification task. In general, the $\mathcal{L}_{orth}^{all}$ maintains an almost across-the-board lead compared to three other independent methods, except for a slightly weaker performance on individual tasks.

Ablation study: The proposed DCLR framework mainly consists of two-stage learning process. The former stage employs the supervised contrastive learning to push the representations of face-voice data from the same person closer while pulling those representations of different person away. The loss function employed for this stage is illustrated in Eq. (1). The latter stage further innovates a multi-layer orthogonal decoupling scheme to learn the disentangled latent representations, and the loss function enrolled for this stage is illustrated in Eq. (15). To investigate the impact of each loss functions, we further design three variant experiments that exclude the loss term from the training process, and adopt w/o to indicate the loss terms excluded during the training process.

Table II shows the performance of different variant experiments. It can be found that the $\mathcal{L}_{sup}$ provides the important foundation for the whole learning framework. That is, the supervised contrastive learning is able to pay more attention to the difficult positive and negative examples in a cross-modal batch, which can well push the representations of face-voice data from the same person closer while pulling those representations of different person away. The $\mathcal{L}_{orth}^{all}$ further promotes the face-voice association performance, and the derived disentangled cross-modal latent representations explicitly provide the shared identity information across the heterogeneous face and voice data. Through the learning of $\mathcal{L}_{orth}^{all}$ and $\mathcal{L}_{rec}$, the interference factors that may bring negative impacts to the cross-modal embedding are filtered into in private embedding, which can enhance the common

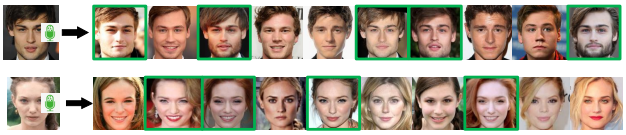
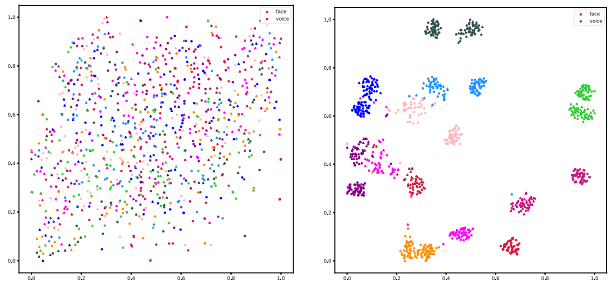


Fig. 5. Top 10 cross-modal retrieval examples, and the correct candidate is marked with a green box.

embedding to benefit various face-voice matching tasks.



(a) Embeddings before training. (b) Embeddings after training.

Fig. 6. Visualization of the learned embeddings.

Parameters Analysis and Visualization: Fig. 4 shows the impact of different dimensions within the proposed DCLR framework. It can be found that the different settings of d_1 and d_2 also can achieve comparable association performance. That is, these parameters are generally insensitive to the face-voice matching performances within a wide range of values.

Further, we show the representative cross-modal face-voice matching examples to visualize the retrieval results. As shown in Fig. 5, it can be observed that the proposed DCLR framework holds a strong ability to associate the semantically similar face and voice examples. This indicates that the proposed disentangled model can well filter the irrelevant information in the cross-modal embedding, and the derived cross-modal latent embedding is valuable to provide significant identity information for discriminative representation.

Besides, we further utilize the t-SNE [23] algorithm to visualize the learned embedding vectors from ten randomly selected peoples. As shown in Fig. 6, it can be found that the proposed DCLR model is able to cluster the face and voice of the same identity close together, while pulling those of different modalities away, and some clusters are tighten closely. Therefore, the disentangled latent representations are semantically meaningful and discriminative for benefiting various face-voice association tasks.

V. Conclusion

This paper proposes an efficient disentangled cross-modal representation learning framework for various face-voice association and matching. The proposed framework first employs the supervised contrastive learning to push the representations of face-voice data from the same person closer while pulling those representations of different person away. Then, the network model further learns the disentangled latent representations that are shared across the face and voice data, while filtering out the modality-dependent factors. Consequently, the proposed method is capable of capture the common semantic information across faces and voices, and the qualitative results have shown its competitive performance.

References

[1] 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.

[2] K. Hoover, S. Chaudhuri, C. Pantofaru, I. Sturdy, and M. Slaney, "Using audio-visual information to understand speaker activity: Tracking active speakers on and off screen," in *Proc. ICASSP*, 2018, pp. 6558–6562.

[3] Y. Wen, M. A. Ismail, W. Liu, B. Raj, and R. Singh, "Disjoint mapping network for cross-modal matching of voices and faces," in *Proc. ICLR*, 2019.

[4] 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 *Proc. SIGIR*, 2020, pp. 1881–1884.

[5] C. Kim, H. V. Shin, T.-H. Oh, A. Kaspar, M. Elgharib, and W. Matusik, "On learning associations of faces and voices," in *Proc. ACCV*, 2018, pp. 276–292.

[6] S. Horiguchi, N. Kanda, and K. Nagamatsu, "Face-voice matching using cross-modal embeddings," in *Proc. ACM MM*, 2018, pp. 1011–1019.

[7] A. Nagrani, S. Albanie, and A. Zisserman, "Seeing voices and hearing faces: Cross-modal biometric matching," in *Proc. CVPR*,2018, pp. 8427–8436.

[8] A. Nagrani, S. Albanie, and A. Zisserman, "Learnable pins: Cross-modal embeddings for person identity," in *Proc. ECCV*, 2018, pp. 71–88.

[9] 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 *Proc. DICTA*, 2019, pp. 1–7.

[10] X. Lin, J. Cao, P. Zhang, C. Zhou, Z. Li, J. Wu, and B. Wang, "Disentangled deep multivariate hawkes process for learning event sequences," in *Proc. ICDM*, 2021, pp. 360–369.

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

[12] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "beta-vae: Learning basic visual concepts with a constrained variational framework,"in *Proc. ICLR*, 2016.

[13] R. Hamaguchi, K. Sakurada, and R. Nakamura, "Rare event detection using disentangled representation learning," in *Proc. CVPR*, 2019, pp. 9327–9335.

[14] 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, 2022.

[15] Z. Wang, B. Sisman, H. Wei, X. L. Dong and S. Ji, "CorDEL: A Contrastive Deep Learning Approach for Entity Linkage," in *Proc. ICDM*, 2020, pp. 1322-1327.

[16] M. Cheng, F. Yuan, Q. Liu, X. Xin, and E. Chen, "Learning transferable user representations with sequential behaviors via contrastive pre-training," in *Proc. ICDM*, 2021, pp. 51–60.

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

[18] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, Article No. 101027, 2020.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition,"in *Proc. BMVC*, 2015, pp. 41.1–41.12.

[20] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI*, 2017.

[21] X. Lu, L. Zhu, Z. Cheng, L. Nie, and H. Zhang, "Online Multi-modal Hashing with Dynamic Query-adaption," in *Proc. SIGIR*, 2019, pp. 715C724.

[22] H. Cui, L. Zhu, J. Li, Y. Yang and L. Nie, "Scalable Deep Hashing for Large-Scale Social Image Retrieval," IEEE Transactions on Image Processing, vol. 29, pp. 1271-1284, 2020.

[23] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, pp. 2579-2605, 2008.