Taking a Part for the Whole: An Archetype-agnostic Framework for Voice-Face Association

Guancheng Chen
Dep. of CS, Huaqiao University &
Zhejiang Lab & Fujian Key Lab. of Big
Data Intelligence and Security
Xiamen, China
guancheng@stu.hqu.edu.cn

Xin Liu*
Dep. of CS, Huaqiao University &
Zhejiang Lab & Xiamen Key Lab. of
Comput. Vis. Pattern Recognit.
Xiamen, China
xliu@hqu.edu.cn

Xing Xu
Center for Future Media & School of
CSE, University of Electronic Science
and Technology of China
Chengdu, China
xing.xu@uestc.edu.cn

Yiu-ming Cheung*
Department of Computer Science,
Hong Kong Baptist University
Hong Kong SAR, China
ymc@comp.hkbu.edu.hk

Taihao Li
Artificial Intelligence Research
Institute of Zhejiang Lab
Hang Zhou, China
lith@zhejianglab.com

ABSTRACT

Voice-face association is generally specialized as a cross-modal cognitive matching problem, and recent attention has been paid on the feasibility of devising the computational mechanisms for recognizing such associations. Existing works are commonly resorting to the combination of contrastive learning and classification-based loss to correlate the heterogeneous datas. Nevertheless, the reliance on typical features of each category, known as archetypes, derived from the combination suffer from the weak invariance of modality-specific features within the same identity, which might induce a cross-modal joint feature space with calibration deviations. To tackle these problems, this paper presents an efficient **A**rchetype-**a**gnostic framework for reliable voice-face association. First, an **A**rchetype-**a**gnostic **S**ubspace **M**erging (**AaSM**) method is carefully designed to perform feature calibration which can well get rid of the archetype dependence to facilitate the mutual perception of datas. Further, an efficient **B**ilateral **C**onnection **R**e-gauging scheme is proposed to quantitatively screen and calibrate the biased datas, namely loose pairs that deviate from joint feature space. Besides, an **I**nstance **E**quilibrium strategy is dynamically derived to optimize the training process on loose data pairs and significantly improve the data utilization. Through the joint exploitation of the above, the proposed framework can well associate the voice-face data to benefit various kinds of cross-modal cognitive tasks. Extensive experiments verify the superiorities of the proposed voice-face association framework and show its competitive performances with the state-of-the-arts.

CCS CONCEPTS

• **Information systems → Novelty in information retrieval**; *Information retrieval*;

KEYWORDS

Voice-Face association, Archetype-agnostic, Re-gauging, Instance Equilibrium

ACM Reference Format:
Guancheng Chen, Xin Liu, Xing Xu, Yiu-ming Cheung, and Taihao Li. 2023. Taking a Part for the Whole: An Archetype-agnostic Framework for Voice-Face Association. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29–November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3581783.3611938

1 INTRODUCTION

Your voice tells strangers what you look like, and it seems to be incredible when you first hear this statement. Generally, people may not realize that such assumptions are based on how they look and sound across different person-specific data samples. In fact, studies in neurosciences have shown that humans have the ability to match the image of an unfamiliar face to an unfamiliar voice with higher accuracy than chance and vice versa, which motivated the machine learning algorithm to emulate the human ability to find the associations between voices and faces intelligently.

Face-voice association can be considered as a cognitive task of finding their semantic correspondence, which is of crucial importance to creating natural human machine interaction systems. The prior works are devoted to exploring the relationship of inter-modal, mining the common attributes between different modal information so that their corresponding features will be comparable in joint space. For instance, Wen *et al.* [35] map different modalities individually by obtaining supervision indirectly from their common covariates, referring to the identity-sensitive factors. Kim *et al.* [16], Nagrani *et al.* [20], Ying *et al.* [6], Wen *et al.*[34], Kai *et al.* [5] directly seek the abstract and entangled connections between different modalities to associate the voice and face data.

In recent years, some works have noted the importance of intra-modal connections that can significantly complement the inter-modal alignment. Along this line, Wang *et al.* [32], Wen *et al.* [34]

[*] Xin Liu and Yiu-ming Cheung are the corresponding authors.

MM '23, October 29–November 3, 2023, Ottawa, ON, Canada.

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3611938

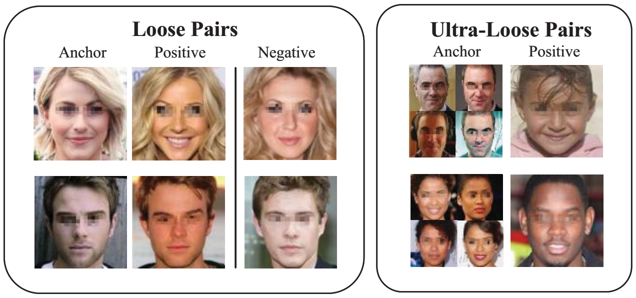


Figure 1: Examples of loose pairs and ultra-loose pairs for face modality. Loose pairs are triplets that fail to meet the specified distance constraint, often due to high similarity in appearance between the positive and negative samples. Differently, the ultra-loose pairs are defined as invalid triplets, for instance, when the anchor and positive samples do not belong to the same identity or contain invalid distorted data.

and Nagrani *et al.* [22] attempt different loss functions to constrain intra-modal representations. Although these works have witnessed the progress of injecting intra-modality constraints into the original model framework, such rough way often ignores the data interaction and is found to be sensitive to the experimental results. Prior works take the ***classification-based*** loss as the classical paradigm of intra-modal representation learning, while the strategy above relies on the global ***archetype***, namely the weight matrix of identity classifier or category center vector. As the invariance of modality-specific representation with consistent identity is the key point of mini-batchrepresentation, such suboptimal indirect manner provided by ***archetype-dependent*** loss will hamper the merge of representative subspaces. Besides, inconsistent handling of different data is necessary in data-driven deep learning models. Wen *et al.* [34] focus more on the hard identities by applying the adaptive weight to the data with consistent identity, while Avishek Joey Bose and Arsha Nagrani *et al.* [15] disregard the limits of identity and impose stronger constraints on hard samples.

Based on the above analysis, the following issues can be summarized: (a) Conventional methods with archetype-dependent loss merely condsider the indirect mutual perception between subspaces, resulting in poor spatial fusion effect. (b) Imposing weights on data with ***category granularity*** are insufficient for handling the loose pairs since there is no guarantee that the quality of data in the same category are the same. (c) Existing methods generally exploit the hard samples on instance granularity, but which intrinsically ignore the hard samples in value and result in excessive mining of hard samples with low value.

For solving problem (a), we propose an efficient **Archetype-agnostic Subspace Merging method** that can directly build a connection among the voice-face representation. For (b) and (c), an efficient **Bilateral Connection re-Gauging** scheme is proposed to quantitatively screen and calibrate the loose pairs that deviate from joint feature space, while an **Instance Equilibrium** strategy is dynamically derived to optimize the training process on loose data pairs and significantly improve the data utilization. In a nutshell, the main contributions of this paper are summarized as follows:

- An efficient **A**rchetype-**a**gnostic **S**ubspace **M**erging strategy (**AaSM**) is explicitly proposed to perform modality alignment across voice-face data, which can well get rid of the archetype dependence to facilitate the mutual perception of datas and promote more accurate feature representations.
- An efficient **Bilateral Connection re-Gauging** scheme is proposed to quantitatively screen and calibrate the loose pairs, which can guide the association framework to learn the semantic correspondence within these hard samples.
- An **Instance Equilibrium** strategy is dynamically derived to optimize the training process on loose data pairs, embed the positive guidance to the learning model and significantly improve the data utilization.
- Extensive experiments evaluated on various face-voice association tasks verify the advantages of the proposed model in comparison with SOTAs.

2 RELATED WORK

2.1 Learning Discriminative Representation

As the first work to realize the cross-modal audio-visual matching task, SVHF [21] concatenate different modality features into one feature, and feed them into a binary classifier. DIMNet [35] believe that covariates such as gender, nationality, etc., which can be extracted from any single modality, can be used as the basis for the mutual mapping of representations between modalities. Although the relatively pure feature information can be obtained from common covariates, the number and attributes of covariates that are difficult to determine limit its performance. Later, VAE-based approaches [17, 30] generate and model the latent space for associating two modalities.

In recent years, the metric learning method has performed well in many fields, e.g. triplet loss, contrastive loss. As confirmed by [3, 12, 13, 31], the effect of metric learning to some extent is proportional to the number of negative pairs. For the sake of further performances, Chen *et al.* [4] propose a quadruplet ranking loss which is modified based on the triplet loss and Horiguchi *et al.* [11, 28] utilize N-pair loss to gain more constrained condition. In [16, 20, 32, 34], the contrastive loss does effectively constrain the inter-modal relationship. What's more, due to the co-occurrence of image and audio in videos, unsupervised audio-visual representation metric learnings [1, 2, 19, 37] can obtain supervisory signal by exploiting spatio-temporal synchronization of information.

Besides, learning of intra-modal associations also improve the discrimination of representations for different identities. In [24, 32, 34], classification-based loss e.g. center loss or identity loss is applied to distinguish intra-modal different identities. It is worth noting that, compared to the contrastive learning which directly acts on the relationship between features, the above-mentioned loss needs to rely on extra reference features as bridges to optimize indirectly. As what have been described above, a main drawback of most existing methods is that indirect intra-modal alignment hinders the fusion of modal common spaces. In addition, such alignment mechanism separation may lead to a gap of the learning rate within and between modalities which hinder the representation learning. Yet in this work, we propose a novel subspace fusion mechanism where archetype representations are not required.

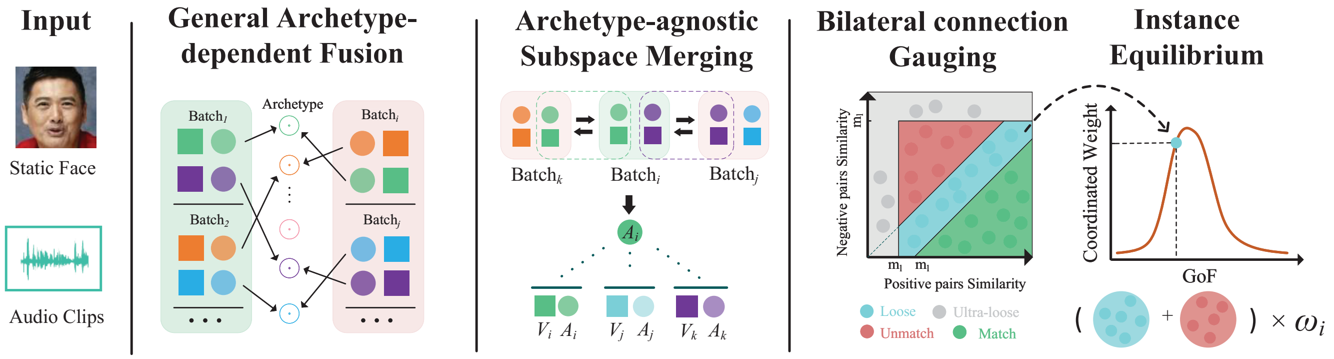


Figure 2: The main differences and comparison with prior works. Previous archetype-dependent works have to rely on extra agent representations, while AaSM promotes the fusion of cross-modal shared subspaces in a archetype-agnostic manner. Further, we re-gauge the bilateral connection of “anchor” quantitatively for sifting loose informative pairs whose value is measured by Instance Equilibrium.

2.2 Hard Sample Mining Method

Samples in datasets often exhibit varying training difficulties. Larger loss function values indicate that the model does not fit certain samples well, and giving more attention to these samples can enhance overall performance. Damianos Galanopoulos *et al.* [9] used a threshold value to exclude potentially-positive samples and identified the hard-negative sample with the highest similarity score. Arsha Nagrani *et al.* [11, 28] employed the "negative difficulty parameter T" to control the difficulty of negative samples obtained from the potential negative list and adopted a curriculum learning strategy for training. Additionally, Florian Schroff *et al.* [27] focused on selecting semi-hard negative exemplars to form appropriate triplets for faster convergence. In works like [14, 25, 33], mining hard samples was used for forming triplets and calculating loss functions.

While the above methods can prioritize hard samples in the model, it’s important to note that hard samples may also contain low-value noise and irrelevant data. These extreme samples prompt us to further analyze and subdivide the hard samples. An effective approach is to assign different weights to individual samples. Wen *et al.* [34] assigned weights to each sample at the identity-level using an adaptive strategy. However, solely relying on the identity-level weighting strategy is inadequate to address instance-level hard samples. To tackle this challenge, we propose a novel strategy to re-gauge and balance the impact of deviated sample pairs on the model at the instance-level.

3 METHODOLOGY

For voice-face association, the most important thing is to map the face and audio embedding into the shared representation space, whereby the comparison can be made across different modalities. Our goal is to learn the face and voice representations that mapped to nearby points of the same identity, while separating the points with different identities.

3.1 Archetype-agnostic Subspace Merging

Prior works [3, 12] have shown the outstanding performance instructed by contrastive learning method. When it comes to multi-modal learning, it is insufficient to learn inter-modal relationships. A combination of contrastive loss and classification-based loss become the workhorse in cross-modal representation learning. Note that, such classified-based loss relies on extra archetype as a mediator and experimental results show that this indirect manner hinders the fusion of mini-batch representation subspaces. So How can we interact directly? Inspired by [36], the Archetype-agnostic Subspace Merging (AaSM) based on the non-parametric softmax is proposed for joint representation Learning.

The traditional methods often utilize the random parameters to build a k-class linear classifier for intra-modal alignment, which includes a $k \times d$-dimensional weight parameter $W^{k \times d}$. Accordingly, the probability for feature $\boldsymbol{x}$ belonging to identity i is calculated:

$$P(i|\boldsymbol{x}) = \frac{\exp(\boldsymbol{x} \cdot \boldsymbol{w}_i)}{\sum_j^B \exp(\boldsymbol{x} \cdot \boldsymbol{w}_j)} \tag{1}$$

where $\boldsymbol{w}_i$ denotes the i-th row of weight matrix $W^{k \times d}$, B indicates the size of the batch. This encourages all features with identity i to be close to weight w_i. In this way, feature aggregation can be treated as feature alignment. However, its shortcoming lies that there is a dependence on the weight w_j , which prevents the mutual perception between features and affects the accuracy of feature alignment. This dependency problem will be solved as follows: Two modality-specific neural networks $f_v(\cdot), f_a(\cdot)$ are utilized to map face images and voices into feature vectors v_i and a_i on a d-dimensional hypersphere. Given a positive feature pair (v_i, a_i) extracted from the same identity i and a negative set $\mathbb{N}^M = \{k_j | j \neq i\}_j^N$ on modality $M \in \{V, F\}$. Mapping representations (v_i, a_i) of the same identity to nearby points while repelling the ones from different identities by minimizing the InfoNCE loss:

$$\mathcal{L}_c(\boldsymbol{a}_i, \boldsymbol{v}_i; \tau) = -\log P(\boldsymbol{a}_i | \boldsymbol{v}_i; \tau) - \log P(\boldsymbol{v}_i | \boldsymbol{a}_i; \tau) \tag{2}$$

$$\begin{aligned} where \quad P(\boldsymbol{y}_i|\boldsymbol{x};\tau) &= \frac{\exp(\boldsymbol{x}\cdot\boldsymbol{y}_i/\tau)}{Z_i} \\ Z_i &= \sum_{\boldsymbol{y}_j\in\mathbb{N}^M}\left\{\exp(\boldsymbol{x}\cdot\boldsymbol{y}_j/\tau)+\left(m\cdot\exp(\boldsymbol{x}\cdot\boldsymbol{y}_i/\tau)\right)\right\} \end{aligned} \tag{3}$$

where $P(\boldsymbol{y}_i|\boldsymbol{x};\tau)$ represents the probability that vector $\boldsymbol{x}$ belongs to the identity i corresponding to vector $\boldsymbol{y}_i$, τ is the temperature hyper-parameter and m is the margin hyper-parameter. Equivalently, the above formula can be simplified to a softmax function. Compared to the softmax function in Eq. (1), the crucial distinction is that feature $\boldsymbol{y}$ in contrastive loss is not derived from the randomly initialized feature archetype, but extracted from the real training data set. This implies that classification-based loss indirectly accomplishes feature classification by guiding the model to collect features onto corresponding archetypes, whereas contrastive loss achieves a similar classification effect directly.

Moreover, the intra-modal alignment task is critical in creating a joint feature space. As the training data is fed in batches, each iteration generates a feature subspace. The goal of intra-modal feature alignment is to align and fuse different features. Therefore, the efficiency of intra-modal alignment tasks significantly impact the formation of the final joint feature space. To this end, we propose an archetype-agnostic subspace merging strategy for fusing feature subspaces through direct mutual perception by injecting the non-parametric mechanism into the within-modal alignment task. Accordingly, the input data is extended to $I_i = \{\boldsymbol{a}_i, \bar{\boldsymbol{a}}_i, \boldsymbol{v}_i, \bar{\boldsymbol{v}}_i \mid \boldsymbol{a}_i \neq \bar{\boldsymbol{a}}_i, \boldsymbol{v}_i \neq \bar{\boldsymbol{v}}_i\}_{i=1}^B$ with auxiliary data $\{\bar{\boldsymbol{a}}_i, \bar{\boldsymbol{v}}_i\}$. We directly utilizes the sample features instead of feature archetypes, enabling mutual perception between features. In terms of voice modality, the weight w_j are replaced by the auxiliary features $\bar{\boldsymbol{a}}_i$, as is the face modality:

$$P(i|\boldsymbol{a}) = \frac{\exp\left(\boldsymbol{a}\cdot\bar{\boldsymbol{a}}_i\right)}{\sum_j^B \exp\left(\boldsymbol{a}\cdot\bar{\boldsymbol{a}}_j\right)} \tag{4}$$

It can be clearly observed that the Eq. (4) and contrastive learning Eq. (2) are linked together. The invariance of intra-modal representation is optimized by minimizing the loss:

$$\mathcal{L}_{\text{wma}} = \sum_i^B \left\{\mathcal{L}_c(\boldsymbol{a}_i, \bar{\boldsymbol{a}}_i;\tau) + \mathcal{L}_c(\boldsymbol{v}_i, \bar{\boldsymbol{v}}_i;\tau)\right\} \tag{5}$$

which gets rid of the archetypes under the classification-based loss. Combining the cross-modal representation alignment loss, the overall Archetype-agnostic Subspace Merging (**AaSM**) loss is formulated as follows:

$$\mathcal{L}_{\text{AaSM}} = \mathcal{L}_{\text{cross}} + \mathcal{L}_{\text{wma}} \tag{6}$$

where $\mathcal{L}_{\text{cross}} = \mathcal{L}_c(\boldsymbol{a}_i, \boldsymbol{v}_i;\tau)$.

Disscussion Classification-based loss functions typically rely on pre-defined global features. For instance, in identity loss-based methods, the weight parameter for identity classification is represented as $W \in \mathbb{R}^{N\times D}$, where N and D denote the number of identities and feature dimensions, respectively. Similarly, in center loss-based methods, the center vector is denoted as $\boldsymbol{c}_{y_i} \in \mathbb{R}^d$, representing the center of deep features for the y_i-th class. Both weight matrix W and center vectors $\boldsymbol{c}_{y_i}$ serve as pre-defined centers for different categories, acting as the archetypes of those categories.

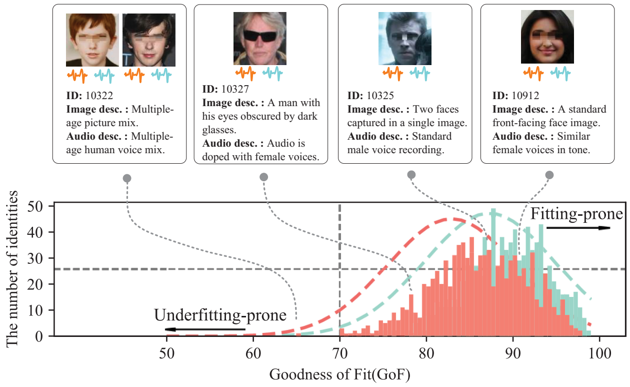


Figure 3: The distribution of GoF and the weights. As the training process progresses, the distribution gradually shifts from red to cyan. Furthermore, the figure illustrates the presence of loose pairs across different identities.

Consequently, alignment tasks between features can be converted into alignment tasks between features and global archetypes. However, the introduction of archetypes increases the training burden as they undergo random initialization and optimization during model training. Additionally, the derived archetypes may hinder direct mutual perception between one another, leading to indirect comparisons with the archetypes as intermediaries. This deviation in modal alignment may result in poor feature matching.

3.2 Bilateral Connection re-Gauging

Although contrastive learning benefits from the large receptive field brought by the extensive set of negative pairs[3, 13, 31], it can only handle the qualitative multilateral connection of the anchor in this scenario.

Next, we refine the model by identifying potential ***loose pairs*** through quantitative re-gauging. Attributive "loose" has two-fold meanings for data pairs: **(1)** Positive pairs with low similarity, such as when the image in the positive pair contains multiple faces or exhibits extremely fuzzy distortion, or the audio includes strong background noise. **(2)** Negative pairs with high similarity. Notably, training with ***ultra-loose pairs*** can adversely impact performance, so we exclude them. While optimizing the loose pairs aids in calibrating the deviated data pairs. Figure 1 shows examples of ultra-loose pairs and loose pairs.

Therefore, we use $\mathbf{M}^u$ and $\mathbf{M}^l$ to filter out the ***ultra-loose pairs*** and ***loose pairs***, respectively. Additionally, we impose more concrete quantization conditions on these pairs. The masks for identifying ultra-loose and loose pairs are formulated as follows:

$$\begin{aligned} \mathbf{M}_i^u &= \mathbb{I}\left\{\left((L * I_n)_{i,i} - m_u\right) < 0\right\} \\ \mathbf{M}_{i,j}^l &= \mathbb{I}\left\{\left(L_{i,j} - (L * I_n)_{i,i} + m_l\right) > 0\right\} \end{aligned} \tag{7}$$

where m_u and m_l is the hyper-parameter threshold. Matrix L and I refers to the inter-modal representations similarity over mini-batch and the unit matrix respectively.

Suppose we capture T_K loose pairs $\{(\boldsymbol{ank}_i, \boldsymbol{pos}_i, \boldsymbol{neg}_j)|i \in K, j \in T_i\}$ from K different identity samples respectively, then define $s_i^p =$

$h(\boldsymbol{ank}_i, \boldsymbol{pos}_i), s_j^n = h(\boldsymbol{ank}_i, \boldsymbol{neg}_j)$ where $h(\cdot)$ is a score function implemented using cosine similarity.

Due to the two-fold causes of loose pairs, its constraint function must be able to adaptively reconcile the two so as to find the correct optimization direction. Inspire by [29], the bilateral relationship re-gauging loss is formulated as follows:

$$\mathcal{L}_{re} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{T_i}\frac{1}{T_i}\left\{\eta_j^n\left(s_j^n - \Delta_n\right) - \eta_i^p\left(s_i^p - \Delta_p\right)\right\} \tag{8}$$

where m is the margin that indirectly controls the between-class $\Delta_p = 1 - m$, within-class margins Δ_n=m, the optimal solution O_p=$(1+m)$, and the coefficient $\eta_j^n = \left[s_j^n - O_n\right]_+$, $\eta_i^p = \left[O_p - s_i^p\right]_+$. The modified loss function using matrix operations is as follows:

$$\mathcal{L}_{re} = \frac{1}{\sum \mathbf{M}}\sum\left\{H^n \odot \left(L - \Delta_n\right) - H^p \odot \left[\left(L \odot I_n\right) * J_n - \Delta_p\right]\right\} \odot \mathbf{M} \tag{9}$$

$$H^p = \left[O_p - L\right]_+ \odot I_n * J_n, H^n{=}\left[L + m\right]_+, \mathbf{M}{=}\mathbf{M}_{i,j}^l \wedge \left(J_n - \mathbf{M}_i^u\right) \tag{10}$$

where H^p, H^n denote the matrix form of coefficients η_i^p, η_j^n in matrix operations, respectively. In particular, the matrix J_n is an all-one matrix.

3.3 Instance Equilibrium

In Figure 3, we observe that the accuracy of data varies widely across different identities. However, high accuracy of identity data does not necessarily mean the absence of loose pairs, as they are still found across all identities. The crucial difference lies in the quantity of loose pairs, which varies among different identities. For easy-to-fit identity data containing very few ultra-loose pairs, focusing on those pairs may hinder the model's ability to accurately fit the remaining data. On the other hand, in cases where the identity data contains a significant proportion of challenging loose pairs, focusing on these pairs can effectively enhance the accuracy of the identity. Therefore, effectively balancing the loose pairs with varying training values is key to optimizing the model's generalization and maximizing the utility of the training data.

To expand this motivation, we establish a mapping curve that relates the degree of fit and weight of distinct identities initially, as in Figure 3. For the sample of a certain identity, the Goodness of Fit (**GoF**) of the sample is approximately proportional to the accuracy of the sample matching during training. Based on the above assumptions, we count the total number of tested items (N^s) and the number of correctly matched items (N^c) from the cosine similarity matrix (L) in each mini-batch. Afterwards, we aggregate this data over T_{update} mini-batches, then the **GoF** of a certain identity n can be approximated as N^c divided by N^s:

$$GoF_n{=}\sum_{t=0}^{T_{update}}\frac{\sum\limits_{i=0}^{B}\left\{\mathbb{I}\left[y\left(L_{i,i}^{(t)}\right) == n\right]\sum\limits_{i\neq j,j=0}^{B}\left(L_{i,j}^{(t)} < L_{i,i}^{(t)}\right)\right\}}{\sum_{k=0}^{T_{update}}\sum_{i=0}^{B}\mathbb{I}\left[y\left(L_{i,i}^{(k)}\right) == n\right] \cdot B} \tag{11}$$

where $y(\cdot)$ represents the mapping from anchor to identity label, T_{update} is a hyperparameter, indicating that every T_{update} iterations will reassign weight coefficient.

The degree of fit is categorized into three sections: fitting-prone, underfitting-prone, and semi-fitting. The weight assigned to the fitting-prone section decreases with increasing degree of fit, while greater weight is assigned to the remaining sections. At first glance, it seems feasible to model the curve simply as an inverse relationship between the weight dependent variable and the GOF independent variable. However, we contend that the intrinsic factor, '***Personalization***', is a crucial determinant of the curve shape. The degree of personalization, defined as the level of specificity of an identity, is inversely proportional to the model's generalization performance. Since a high degree of ***personalization*** will lead to the underfitting-prone samples, so identity categories with high personalization and low GoF are also assigned small weights.

Further, the Instance Equilibrium strategy is utilized to suppress the fitting-prone and underfitting-prone data while focusing more on semi-fitting data. After training the model for a specified number of iterations, we display the distribution of training data, as shown in Figure 3. Since the bell curve of Gaussian distribution coincides with our weight distribution strategy, a slight adjustment is made to the probability density function (PDF) of Gaussian distribution in order to regulate the weight curve:

$$\boldsymbol{\omega}_i = G\left(GoF_i; \mu + \lambda\sigma, -\xi\sigma^2\right) \tag{12}$$

where $G(\cdot)$ is the probability density function of the Gaussian distribution $N\left(\mu + \lambda\sigma, \xi\sigma^2\right)$, μ and σ denote the mean and standard deviation, respectively. To improve the accuracy of model matching, it is necessary to apply a left offset to the curve for fine-tuning. The hyperparameters λ and ξ are used to control the shape of the curve which sensitive to our strategies. Negative hyperparameters λ is used to implement the left offset, namely $|\lambda\sigma|$. Imposing the weights calculated by Eq. (12) on the re-gauging loss, then it can be modified as:

$$\mathcal{L}_{re} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j=1}^{T_i}\frac{\boldsymbol{\omega}_i}{T_i}\left\{\eta_j^n\left(s_j^n - \Delta_n\right) - \eta_i^p\left(s_i^p - \Delta_p\right)\right\} \tag{13}$$

We then minimize the weighted sum of the two losses:

$$\mathcal{L}_{all} = \mathcal{L}_{\text{AaSM}} + \mathcal{L}_{re} \tag{14}$$

Apply to loose pairs *vs.*all pairs: Because of the radical strategy of instance equilibrium, some identities are given zero weight. Base on this premise, when imposing weight on all loss functions, the model can only passively fit identity samples with zero weights according to the patterns of other identities. What's worse, due to the instability of the initial weight, the proportion of zero-weight samples will be increased which will keep a lid on model performance. Therefore, instead of covering all the loss functions with weight, we feed all the data indiscriminately into the Bilateral connection re-Gauging loss to ensure the stability of the model training.

4 EXPERIMENTS

4.1 Datasets

We evaluating our method on two datasets including combined dataset Vox-CM which still face images and audio flips are extracted from availabled dataset VGGFace[26] and VoxCeleb[23]

Table 1: Details of the datasets.

Dataset	Partition	Train	Validation	Test	Total
Vox-CM	Voice clips	113,322	14,182	21,850	12,362
	Face images	104,724	12,260	20,076	91,238
	Identities	924	112	189	1,225
	Queries(V2F)	–	42,546	65,550	108,096
	Queries(F2V)	–	36,780	60,228	97,008
AVSpeech-CM	Voice clips	12,357	1,505	3,096	16,958
	Face images	91,238	8,631	18,045	117,914
	Identities	400	50	100	550
	Queries(V2F)	–	43,645	54,135	97,780
	Queries(F2V)	–	43,155	52,632	95,787

respectively and a new dataset AVSpeech-CM constructed in-house which will be made publicly available soon.

Vox-CM VGGFace comprises of 982,803 still face images with 2.6M identities collected by querying raw images in Image Search. Meanwhile, VoxCeleb is an audio-visual dataset consisting of short clips of human speech, containing over 100,000 utterances for 1,251 celebrities with video and audio format. To meet the requirement of proposed tasks, we use only data with 1,225 identities overlapped between VGGFace and VoxCeleb. For fair comparison, we follow the train/val/test split strategy from Wen *et al.*[34] in our experiments. Note that no intersection among these sets.

AVSpeech-CM We introduce a new audio-visual dataset for cross-modal face and voice association learning base on large-scale dataset AVSpeech [8] which comprising speech clips with no interfering background signals. Since in each clip the only visible face and audible sound belong to a single speaking person, such inherent mapping can be broken down to get the data pair we want. Specifically, to accommodate the needs of various tasks of cross-modal face and voice association learning, the dataset creation pipeline is proposed as follows: First, we separate the video frames and audio from the video. Secondly, face images are obtained from the video frames by the OpenCV face recognition method while audio clips are divided according to AVSpeech author annotations. Third, we pre-train a gender prediction model on VoxCeleb and apply it on AVSpeech to obtain gender labels while filtering the data with low confidence level. Based on the criterion of gender balance and high quality of individual data, we screen out 550 identities to make up our dataset, named as AVSpeech-CM (Cross-Modal). The train/validation/test sets is divided according to the ratio of 8:1:2.

To reduce the impact of random data, the tasks of validation and test are based on the queries lists provided in advance which specify the data pairs of voices and faces. More statistical information for the datasets are shown in the Table 1.

4.2 Implementation Details

Network Architecture The architecture of our network is comprised of two sub-networks which are utilized to extract features: the FaceSubnetwork and the VoiceSubnetwork, which are fed with still images of faces and audio clips repectively.

FaceSubnetwork. The faceSubnetwork is implemented with ResNet 152. The size of the image inputted is 112,112,3, which is normalized to [-1, 1] by subtracting 127.5 and dividing 127.5. The output feature is 128 dimensions to which is remapped by the followed fully connected layer from 512 dimensions.

VoiceSubnetwork. The voiceSubnetwork is implemented with ThinResNet34, of which inputs is a spectrogram of audio. Briefly, the MelSpectrogram utilized as input is generated using mono raw audio signal by applying Short-time Fourier Transform STFT and triangular filters with 40 mel filterbanks and a hamming sliding window of width 25ms and of hop 10ms.

Training Strategy. The training process can be divided into two phases: warming up training and training with coordinated weight coefficients. The weight of best performance on validation set will be saved for evaluation. To speed up model convergence, we sample different identities in each mini-batch to ensure maximum identity diversity. The faceSubnetwork is pre-trained on MS-1M [10] driven by face recognition task while voiceSubnetwork is pre-trained on VoxCeleb2 [7] driven by audio speaker recognition.

For data-processing, we apply data augmentation for images, *e.g.*random cropping, flipping and apply random audio duration clipping for audios. Meanwhile, we adopt stochastic gradient descent (SGD) optimizer with mini-batch size of 20 and 0.9 momentum. The learning rate is initialized as $1e-2$, and decays by 0.1 in the 6888 and 9888 iterations. Hyper-parameters set in our model are as follows: $m = 3.4$, $m_l = 0.2$, $m_u = 0.15$, $T_{warm} = 6440$, $T_{update} = 920$.

Evaluation protocol To assess the performance of model, the following evaluation tasks are selected.

(a) **1:2 matching** Given three samples consist of one face image as probe and two audio segment as gallery, or vice verse. There is only one sample, known as positive candidate, in the gallery whose identity is consistent with query. While the rest, of course, are called negative candidates. The task is to find out candidates in the gallery who match the identity of query. The performance is measured with accuracy (ACC).

(b) **1:N matching** The task is essentially the same as 1:2 matching, and the only change is to enlarge the number N of negative candidates. We restricted N to the range of 2 to 10, and the accuracy is utilized to measure the performance.

(c) **Verification** Given two instances sampled from different modalities. The task is to determine whether they belong to the same identity or not. The performance is measured with Area Under the ROC curve (AUC).

(d) **Retrieval** This task is more challenging than the 1: N matching task because it has more than one positive candidates in the gallery. This task is to rank the gallery to make sure those with high similarity to probe are ranked at the top. The performance is measured with mean average precision (mAP) [18].

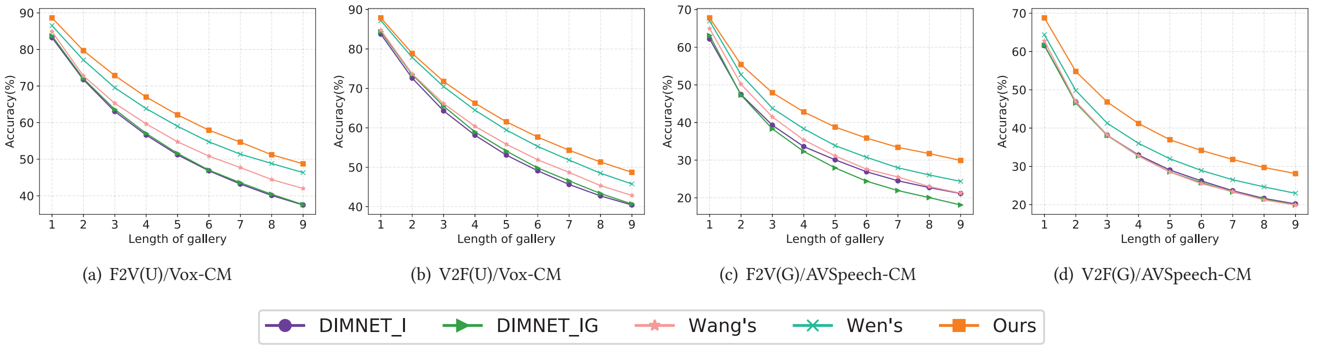
We obtain the measurement of each task in two different scenarios, voice to face (V2F) and face to voice (F2V). Similarly, we subdivide the query lists for testing into two categories: gender unrestricted (**U**) and gender restricted (**G**). This means that when testing in the gender restricted list, the model cannot regard gender as the basis for judgment.

4.3 Quantitative Results

We compare our proposed method with four representative models, including SVHF[21], DIMnet[35], Wang's[32], Wen's[34]. Due to the limitation of the model structure, only partial task results are

Table 2: The result of multiple tasks on Vox-CM and AVSpeech-CM. V2F:from voice to face, F2V:from face to voice; U: gender unrestricted; G: gender restricted. The best results of our models are shown in bold.

Datasets	Tasks	1 : 2Matching (ACC)				Verification (AUC)				Retrieval (mAP)	
		V2F(U)	F2V(U)	V2F(G)	F2V(G)	V2F(U)	F2V(U)	V2F(G)	F2V(G)	V2F	F2V
Vox-CM	SVHF[21]	81.00	79.50	63.90	63.40	–	–	–	–	–	–
	DIMNet-I[35]	84.06	82.92	73.32	70.29	82.69	82.61	71.51	69.81	5.50	3.92
	DIMNet-IG[35]	84.39	83.70	73.15	71.67	83.10	83.56	72.07	71.14	4.68	4.01
	Wang's[32]	84.76	84.87	74.10	74.22	84.25	84.87	74.63	74.74	5.13	4.45
	Wen's[34]	87.20	86.50	77.70	75.30	87.20	87.00	75.50	76.10	5.50	5.80
	Ours	**87.87**	**88.63**	**78.90**	**78.58**	**88.42**	**89.02**	**79.17**	**78.95**	**6.77**	**7.16**
AVSpeech-CM	SVHF[21]	67.28	69.96	51.25	54.03	–	–	–	–	–	–
	DIMNet-I[35]	70.07	70.68	61.53	62.18	68.8	70.36	60.42	60.84	7.23	9.07
	DIMNet-IG[35]	70.49	71.01	61.63	63.11	70.2	71.53	60.54	61.53	8.68	5.52
	Wang's[32]	72.04	72.54	62.75	64.95	72.56	74.67	62.74	65.03	7.01	5.99
	Wen's[34]	72.53	74.36	64.38	66.96	73.08	74.58	64.07	65.86	9.55	8.49
	Ours	**80.29**	**79.98**	**68.78**	**67.81**	**80.69**	**81.16**	**68.63**	**69.05**	**11.80**	**13.43**



(a) F2V(U)/Vox-CM (b) V2F(U)/Vox-CM (c) F2V(G)/AVSpeech-CM (d) V2F(G)/AVSpeech-CM

Figure 4: Quantitative results on 1:N matching task.

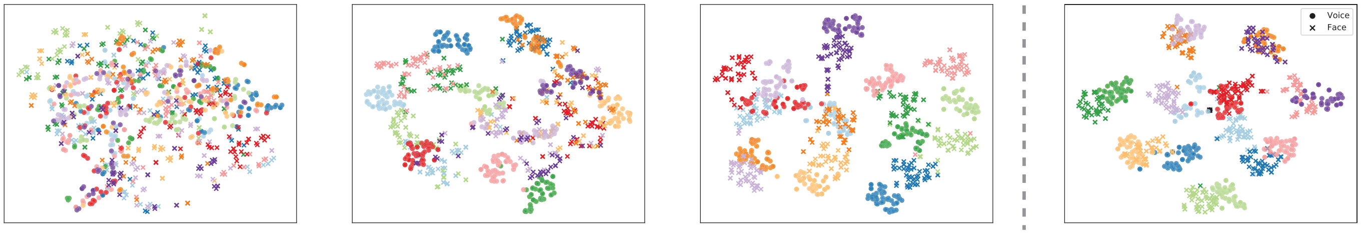
available. Table 2 shows the quantitative results of our method and its competitors on both datasets. The performance of our method on all tasks significantly surpass all the competitors. Compared to the state-of-the-art on Vox-CM, except for the small improvement on task 1 (about 0.67%), other tasks have obvious improvement, bringing an average performance gain of 2%.

In the horizontal comparison of Table 2, it is obvious that the presence or absence of gender constraints has a great influence on the model results. The results of model drop by an average of 9% with the same gender constraint on the training set. This suggests that gender plays a key role in inter-modal associations. Compared to previous methods, it is clear that this gap is narrowed in our method. Especially on the validation task of V2F and the matching task of F2V without gender restriction, performance gains as high as 3.67% and 3.28% are achieved, respectively. Such large improvement has been obtained with the absence of the discriminant factor of gender fully demonstrates that our method can mine deeper gender-irrelevant connections.

To further verify the generalization performance of our method, we evaluate our method on AVSpeech-CM and report the corresponding results in Table 2. Similar conclusions can be obtained for AVSpeech-CM as for Vox-CM. What's more, we show the results of the proposed method on the 1:N matching task on two datasets in Fig 4. As the value of N increases, the difficulty of the matching task keeps increasing and the gap between the previous method and our method keeps getting bigger. On the other hand, the performance of the previous method drops significantly, while our method drops relatively gently. It shows that the features extracted by our method are more accurate and discriminative.

4.4 Qualitative Results

We used two different markers and various colors to distinguish between modal information and identity, respectively. We then visualized the feature extraction results of the model under different iterations, as shown in Figure 5. The figure illustrates that as the model iteratively trains, features of the same identity consistently cluster together. Furthermore, we observed that audio information clusters together before image information. This could be attributed to audio's smaller information capacity and the model's faster fitting degree. The image on the far right demonstrates that the model effectively clusters features of the same identity together under optimal weight, and also establishes connections between different



(a) Chance (b) After 600 iterations (c) After training (d) Test set visualization

Figure 5: Visualization results of voice and face representations on Vox-CM. Distinct identities are denoted by different colors, while dissimilar markers are used to represent embeddings of different modalities.

modal features. That's means the learned representations cross the boundaries of modalities and be mapped into a shared space, demonstrating the effectiveness of the proposed method.

However, we noticed clusters of different modal information of different identities in the figure, which could be due to the fact that two-dimensional data points after dimensionality reduction could not accurately express the distance relation of high-dimensional features. From another point of view, this observation also proved the existence of different identities with high matching similarity in the test set.

4.5 Ablation Study

We perform an ablation study to demonstrate the effectiveness of the proposed three module: 1) the archetype-agnostic subspace merging, 2) the bilateral connection re-gauging and 3) the instance equilibrium. For 1), we build the model with archetype-dependent loss strategy as opposed to our archetype-agnostic strategy with other modules are not modified. Specifically, we use identity loss and center loss to implement two different archetype-agnostic loss respectively. For 2) and 3), We remove the relevant modules and leave the remaining modules unchanged. We show the results on data set Vox-CM in Table 3.

Effectiveness of the archetype-agnostic subspace merging. Comparing the subspace fusion driven by archetype-agnostic loss to the ones driven by archetype-dependent loss which implemented by center loss and identity loss is marked with (a) and (b), respectively. We observe that AaSM obtains obvious improvement (e.g. on Vox-CM, AaSM brings about 1.10% to 4.30% performance improvement). This is because although the archetype-dependent loss function makes the sample instances tend to the corresponding archetype, it does not mean different samples can be in a reasonable position in the feature space. This comparison demonstrates the AaSM can act directly on the sample relationship so that the mutual perception between samples can ensure that the sample representations are more accurate which promotes the fusion of the modal subspace.

Effectiveness of the bilateral connection re-gauging. Comparing (d) and (e), we can deduce that the module of bilateral connection re-gauging bring about 1.70% improvement which manifests its effectiveness. Such improvements can be explained by deeper exploration: Removing the re-gauging module will result in the unavailability of the loose pairs, thus degrading instance equilibrium from instance-level to identity-level. Intuitively, instance equilibrium mechanism will impose the weights on all the instances. It means that even the non-loose pairs that are easy to learn can be given heavy weight. That's why filtering out the loose pairs by the bilateral connection re-gauging can work.

Table 3: Ablation studies of different modules.

Archetype		Bilateral	Identities	Match.	Verif.
-agnostic	-depend.	re-Gauging	equilib.		
	✔	✔	✔	(a) 80.3	81.0
	✔	✔	✔	(b) 86.2	86.1
✔		✔	✘	(c) 87.3	87.8
✔		✘	✔	(d) 87.1	87.2
✔		✔	✔	(e) **88.6**	**89.0**

Effectiveness of the instance equilibrium. Comparing (c) and (e), it can be seen after imposing the instance equilibrium strategy, the model gain an average increase of about 1.41% which confirms the idea that loose pairs have different values. Hence, the instance equilibrium is indispensable to loose pairs.

5 CONCLUSION

This paper has proposed a novel archetype-agnostic framework for learning the association between voice and face, which consists of the following three modules: 1) An archetype-agnostic subspace merging strategy is utilized to perform feature calibration which can well get rid of the archetype dependence. 2) An efficient Bilateral Connection re-Gauging scheme is proposed to quantitatively screen the loose pairs, featuring on calibrating the biased data pairs. 3) The instance equilibrium strategy is proposed for focusing more on valuable loose pairs while suppress the counter-productive effect of ultra-loose pairs. The experiment results have shown that the proposed voice-face association learning framework has shown its outstanding and improved performance.

ACKNOWLEDGMENTS

This work was supported in part by the Open Project of Zhejiang Lab under Grant 2021KH0AB01, in part by NSFC/Research Grants Council (RGC) Joint Research Scheme under Grant N_HKBU214/21, in part by General Research Fund of RGC under Grant 12201321 and 12202622, and in part by the National Science Foundation of Fujian Province under Grant 2020J01083 and Grant 2020J01084.

REFERENCES

[1] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. 2020. Self-supervised multimodal versatile networks. *Advances in Neural Information Processing Systems* (2020), 25–37.

[2] Guangyu Chen, Deyuan Zhang, Tao Liu, and Xiaoyong Du. 2022. Self-Lifting: A Novel Framework for Unsupervised Voice-Face Association Learning. In *Proceedings of the International Conference on Multimedia Retrieval (ICML)*. 527–535.

[3] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the International Conference on Machine Learning (ICML)*. 1597–1607.

[4] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 403–412.

[5] Kai Cheng, Xin Liu, Yiu-ming Cheung, Rui Wang, Xing Xu, and Bineng Zhong. 2020. 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 (ACM MM)*. 448–455.

[6] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. 2020. Look, Listen, and Attend: Co-Attention Network for Self-Supervised Audio-Visual Representation Learning. In *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*. 3884–3892.

[7] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. 2018. VoxCeleb2: Deep Speaker Recognition. In *Int. Conf. on Interspeech*.

[8] Ariel Ephrat, Inbar Mosseri, Oran Lang, et al. 2018. Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation. In *ACM Transactions on Graphics*. Article 112, 11 pages.

[9] Damianos Galanopoulos, Mezaris, et al. 2021. Hard-Negatives or Non-Negatives? A Hard-Negative Selection Strategy for Cross-Modal Retrieval Using the Improved Marginal Ranking Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*. 2312–2316.

[10] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *Proceedings of the conference on European Conference on Computer Vision (ECCV)*. 87–102.

[11] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 297–304.

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2019. Momentum Contrast for Unsupervised Visual Representation Learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition (CVPR)*. 9729–9738.

[13] Olivier Henaff. 2020. Data-efficient image recognition with contrastive predictive coding. In *Proceedings of the International conference on machine learning (ICML)*. 4182–4192.

[14] Alexander Hermans, Lucas Beyer, and Bastian Leibe. 2017. In Defense of the Triplet Loss for Person Re-Identification. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*.

[15] Miyuki Kamachi, Harold Hill, Karen Lander, and Eric Vatikiotis-Bateson. 2003. Putting the face to the voice': Matching identity across modality. *Current Biology* 13, 19 (2003), 1709–1714.

[16] Changil Kim, Hijung Valentina Shin, Tae-Hyun Oh, et al. 2019. On Learning Associations of Faces and Voices. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*. 276–292.

[17] Minyoung Kim, Ricardo Guerrero, and Vladimir Pavlovic. 2020. Learning Disentangled Latent Factors from Paired Data in Cross-Modal Retrieval: An Implicit Identifiable VAE Approach. *arXiv preprint arXiv:2012.00682* (2020).

[18] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. 2021. MTFH: A Matrix Tri-Factorization Hashing Framework for Efficient Cross-Modal Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (2021), 964–981.

[19] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. 2021. Audio-visual instance discrimination with cross-modal agreement. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 12475–12486.

[20] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Learnable PINs: Cross-modal Embeddings for Person Identity. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 73–89.

[21] Arsha Nagrani, Samuel Albanie, and Andrew Zisserman. 2018. Seeing Voices and Hearing Faces: Cross-modal biometric matching. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 8427–8436.

[22] Arsha Nagrani, Joon Son Chung, Samuel Albanie, et al. 2020. Disentangled speech embeddings using cross-modal self-supervision. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6829–6833.

[23] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. VoxCeleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612* (2017).

[24] Shah Nawaz, Muhammad Kamran Janjua, Ignazio Gallo, Arif Mahmood, and Alessandro Calefati. 2019. Deep latent space learning for cross-modal mapping of audio and visual signals. In *Proceedings of the conference on Digital Image Computing: Techniques and Applications (DICTA)*. 1–7.

[25] Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. 2016. Deep Metric Learning via Lifted Structured Feature Embedding. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 4004–4012.

[26] Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2015. Deep Face Recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

[27] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. FaceNet: A Unified Embedding for Face Recognition and Clustering. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

[28] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class N-pair loss objective. *Advances in Neural Information Processing Systems (NIPS)* 29 (2016).

[29] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. 2020. Circle Loss: A Unified Perspective of Pair Similarity Optimization. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 6398–6407.

[30] Thomas Theodoridis, Theocharis Chatzis, Vassilios Solachidis, et al. 2020. Cross-modal Variational Alignment of Latent Spaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 960–961.

[31] Yonglong Tian, Dilip Krishnan, and Phillip Isola. 2020. Contrastive multiview coding. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 776–794.

[32] Rui Wang, Xin Liu, Yiu-ming Cheung, Kai Cheng, Nannan Wang, and Wentao Fan. 2020. 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 (SIGIR)*. 1881–1884.

[33] Xiaolong Wang and Abhinav Gupta. 2015. Unsupervised Learning of Visual Representations Using Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2794–2802.

[34] Peisong Wen, Qianqian Xu, Yangbangyan Jiang, et al. 2021. Seeking the Shape of Sound: An Adaptive Framework for Learning Voice-Face Association. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 16347–16356.

[35] Yandong Wen, Mahmoud Al Ismail, Weiyang Liu, et al. 2018. Disjoint Mapping Network for Cross-modal Matching of Voices and Faces. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[36] Zhirong Wu, Yuanjun Xiong, Stella X Yu, et al. 2018. Unsupervised Feature Learning via Non-parametric Instance Discrimination. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*. 3733–3742.

[37] Boqing Zhu, Kele Xu, Changjian Wang, Zheng Qin, Tao Sun, Huaimin Wang, and Yuxing Peng. 2022. Unsupervised Voice-Face Representation Learning by Cross-Modal Prototype Contrast. *arXiv preprint arXiv:2204.14057* (2022).