计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202220411
60(11): 2638−2649, 2023

基于双向伪标签自监督学习的跨人脸-语音匹配方法

朱明航[1,2] 柳 欣[1,3] 于镇宁[1,2] 徐 行[4] 郑书凯[3]

[1](华侨大学计算机科学与技术学院 福建厦门 361021)

[2](福建省大数据智能与安全重点实验室(华侨大学) 福建厦门 361021)

[3](之江实验室 杭州 311121)

[4](电子科技大学计算机科学与工程学院 成都 611731)

(mhzhu@stu.hqu.edu.cn)

Cross Face-Voice Matching Method via Bi-Pseudo Label Based Self-Supervised Learning

Zhu Minghang[1,2], Liu Xin[1,3], Yu Zhenning[1,2], Xu Xing[4], and Zheng Shukai[3]

[1](*College of Computer Science and Technology*, *Huaqiao University*, *Xiamen*, *Fujian* 361021)

[2](*Fujian Key Laboratory of Big Data Intelligence and Security*(*Huaqiao University*), *Xiamen*, *Fujian* 361021)

[3](*Zhejiang Lab*, *Hangzhou* 311121)

[4](*School of Computer Science and Engineering*, *University of Electronic Science and Technology of China*, *Chengdu* 611731)

Abstract Neurocognitive science research shows that human brain often combines face information on cross-modal interaction analysis during the speech perception process. Nevertheless, existing cross-modal face-voice association methods still face the various challenges such as sensitivity to complex samples, lack of supervised information and insufficient semantic correlation, which mainly due to the lack of mining common semantic embeddings. To tackle these problems, we present an efficient cross-modal face-voice matching method from bi-pseudo label based self-supervised learning. First of all, we introduce a cross-modal weighted residual network to learn face-voice common embeddings, and then propose a novel self-supervised learning method for bi-pseudo label association, which learns the latent semantic supervision of one modality to supervise the feature learning of another modality. Accordingly, based on this interactive cross-modal self-supervised learning, the highly correlated face-voice associations can be well learned. Besides, in order to increase the discrimination of mining supervised information, we further construct two auxiliary losses to make the face-voice features of the same samples closer, while pushing the features of different samples to be far away. After a large number of experiments, the innovative method proposed in this paper has achieved a comprehensive improvement in the cross-modal face-voice matching task compared with the existing work.

Key words face-voice association; bi-pseudo label; self-supervised learning; weighted residual network; latent semantic supervision

收稿日期: 2022-05-21; **修回日期**: 2022-09-09

基金项目: 之江实验室开放课题 (2021KH0AB01); 国家自然科学基金项目 (61673185, 61976049); 福建省自然科学基金项目 (2020J01083, 2020J01084); 福建省本科高校教学研究项目(FBJG20220092)

This work was supported by the Open Research Projects of Zhejiang Lab (2021KH0AB01), the National Natural Science Foundation of China (61673185, 61976049), the Natural Science Foundation of Fujian Province (2020J01083, 2020J01084), and the Undergraduate Education and Teaching Research Project of Fujian University (FBJG20220092).

通信作者: 柳欣(xliu@hqu.edu.cn)

摘　要　神经认知科学研究表明，人类大脑在感知语音的过程中常常将结合人脸信息进行跨模态交互分析. 然而，现有的跨模态人脸-语音关联方法仍面临着对复杂样本敏感、监督信息缺乏以及语义关联不足等挑战，其主要原因是缺少对潜在共性语义的挖掘. 针对这些问题，提出了基于双向伪标签自监督学习的跨模态学习架构，用于人脸-语音关联学习与匹配任务. 首先，构建跨模态加权残差网络来学习人脸-语音的跨模态共享嵌入，然后提出一种新颖的双向伪标签关联的自监督学习方法，旨在通过一种模态的潜在语义信息去监督另一个模态的特征学习，从而基于这种交互式跨模态自监督学习能够挖掘到人脸-语音间更紧密的关联. 为增加挖掘监督信息的判别性，进一步构建了 2 个辅助损失促使来自相同身份的人脸-语音特征更接近，并使来自不同身份的特征更加疏远. 基于大量实验验证，相比较于现有方法，在人脸-语音跨模态匹配任务上获得了全面的提升.

关键词　人脸-语音关联；双向伪标签；自监督学习；加权残差网络；潜语义监督

中图法分类号　TP391

神经认知学研究表明人类在视听信息感知的时候具有将人脸和语音进行关联的能力. 例如，当人们在跟朋友打电话时，虽然只能听见对方的声音，但是脑海中会下意识地构建出他们的样貌特征，以及当看到好友的照片时，能瞬间回忆起对方熟悉的声音. 心理学研究将人类这种跨人脸-语音的交互能力称为“麦格克效应”[1]，其表明人在跟外界交谈时，能够在人脸信息和语音信息之间进行交互关联. 同时，大量神经科学研究表明，人类的大脑存在着多个模块感知区域[2]，这些区域之间通过脑神经连接来并行处理信息，其中视觉模块跟听觉模块的神经连接通路更为紧密. 基于此，人们可以较好地对人脸与语音进行跨模态感知，例如当人们在观看有声电视节目时，能以较高的准确率从当前说话人声音中匹配出正确的人物身份. 在跨模态生物特征感知方面，人脸信息跟语音信息都能有效地作为刻画人物特定语义的特征载体，比如说身份、性别、年龄、种族和国籍等，这些反映特定语义的生物特征信息可以隐式地从人脸或语音信息中进行提取，从而基于人脸和语音的跨模态交互关联研究具有一定的可行性.

上述现象表明，人脸信息与语音信息之间存在显著的关联特性. 因此，有效的人脸-语音相关性挖掘和跨模态匹配研究能够促进认知科学和人工智能技术创新实践的发展，具有重要的现实意义. 受此启发，越来越多的研究者认识到探索人脸-语音关联的研究已迫在眉睫. 该研究有广阔的应用前景，例如基于语音视频的说话人身份标注、视频人脸及语音信息时态同步和基于声音的人脸面部特征还原等[3-5]. 从国内外研究进展分析，目前的跨人脸-语音模态的研究还依然处于起步阶段[6]，大量人脸-语音关联语义的研究等待着人们去探索.

跨人脸-语音模态关联性学习方法的研究作为一项新颖的课题，存在着许多具有挑战性的任务. 一些方法[7]虽然也对人脸-语音进行了关联性学习，但在跨模态匹配任务的表现中只取得了比随机概率略好的性能表现. 根据现有的人脸-语音关联学习方法，目前跨人脸-语音模态的研究依然面临着 3 个主要挑战：1）样本复杂性，人脸样本和语音样本分别通过不同的传感器获取，它们的特征属性及数据类型完全不同，因此无法直接进行人脸-语音特征间的交互关联，从而导致语义表征间存在着巨大的语义鸿沟. 2）监督信息匮乏，基于有限的标签信息去监督人脸-语音特征，挖掘出的跨模态特征表示过度依赖于人为监督，导致获取的跨模态连接并不可靠，从而无法得到模态间紧密的语义关联. 同时基于有监督的标签生成需要人工的手动注释，其过程繁琐且成本高昂. 3）语义关联不足，现有的大多数人脸-语音跨模态关联方法只是利用损失函数进行简单的特征关联，其从本质上忽略了人脸-语音模态间潜在语义的关联特性，且无法满足实际应用的需求.

针对上述挑战，设计一种可以利用潜在语义促进跨人脸-语音模态关联性学习的方法尤为重要. 值得注意的是，自监督学习旨在通过对原始数据特征中潜在语义的挖掘生成伪标签进而监督整体特征学习，这种从数据本身出发学习特征表示的方法为跨模态关联学习提供了借鉴意义.

基于自监督学习对潜在语义挖掘的思想，本文提出了一种基于双向伪标签自监督学习的跨人脸-语音匹配方法（cross face-voice matching method via bi-pseudo label based self-supervised learning, Bi-Pcm），用于跨模态下的人脸-语音关联与匹配. 具体来说，首先，设计了一个跨模态加权残差网络（cross-modal

weighted residual network, CMWR)模块，在解决人脸-语音特征异构性的同时，学习到模态间的共享嵌入特征. 接着引入自监督学习模块，通过人脸和语音特征间的潜在语义生成伪标签，实现跨模态下的双向监督，进而获取潜在语义关联. 然后，本文对2种模态下生成的伪标签构建关联损失，约束伪标签生成，从而获取基于潜在语义生成的强跨模态嵌入. 最后，通过本文方法获取的跨模态表示将在所有人脸-语音跨模态匹配任务上进行测评. 本文的主要贡献包括4点：

1) 提出了一种新颖的基于双向伪标签自监督学习的方法用于获取人脸-语音间的跨模态关联. 据文献[6−7]所知，本文提出的方法是利用伪标签来促进跨人脸-语音模态下的关联性学习.

2) 设计了一种高效的伪标签生成方法，旨在利用特征空间的关联促进潜在语义对齐，增强相同人脸-语音特征相关性，并扩大不相关人脸-语音之间的特征差异，从而生成高质量伪标签进行监督约束.

3) 创新性地提出了一种基于自监督跨模态学习框架来获取人脸-语音间的共享特征嵌入，并通过一种模态的伪标签语义作为监督信号来监督另一种模态的特征学习，从而高效地进行跨模态语义关联.

4) 大量实验结果表明，本文方法相比较于现有的跨人脸-语音匹配工作，可扩展性更强，并在多个跨人脸-语音模态匹配任务上都取得了全面的提升.

1 相关工作

人类面部视觉及语音信息是人机交互过程中最为直接和灵活的方式，因此基于人脸和语音的跨模态感知吸引了研究学者的广泛关注. 从生物特征角度来看，来自相同身份的人脸和语音数据，对应着许多相似的语义特征，例如性别、种族还有年龄[8]，因此人脸和语音具有表征相同身份的语义关联信息. 目前基于人脸-语音关联特征的方法主要分为2类：基于分类损失和基于空间距离度量. 基于分类损失的代表方法是SVHF[9](seeing voices and hearing faces: cross-modal biometric matching)，它利用卷积神经网络(convolutional neural network, CNN)架构学习人脸-语音间的关联表示，进而解决跨模态匹配任务. 基于空间距离度量的代表方法是PINs[10](learnable pins: cross-modal embeddings for person identity)，该方法通过获取人脸图片和语音片段构建正负例人脸-语音样本对，然后构造个人身份节点，利用对比损失最小化正例样本的空间距离来学习人脸-语音间的嵌入特征. 上述2种方法在一些具有挑战性的实验中，可以达到与人类相当的水平，但是却拥有局限性，即它们所学习出来的特征只能运用于特定的跨模态匹配任务上，当任务更改时网络也需要重新训练.

随着跨人脸-语音模态关联研究的发展，设计能用于多个跨模态匹配任务的通用特征表示引起注意. 在FV-CME[11](face-voice matching using cross-modal embeddings)中首先利用2个分支网络来分别学习人脸和语音模态下的特征表示，并利用N对损失来规范特征对应. 这种方法虽然可以运用于多种人脸-语音的跨模态匹配任务，但需要大量的参数用于模型的优化. LAFV[12](on learning associations of faces and voices)利用对人脸-语音公共信息的整合，学习交叉模态下的特征关联，从而减少跨模态差异，且可以达到与文献[7,13]中方法相似的结果. DIMNet[5](disjoint mapping network for cross-modal matching of voices and faces)使用不相交映射网络(disjoint mapping network)将关联特征映射到共享协变量中，实现了人脸-语音匹配任务上的性能提升. 然而这种学习需要对大规模训练数据进行标签注释，过程耗时且成本昂贵. 为避免使用三元组损失[14]，SSNet[3](deep latent space learning for cross-modal mapping of audio and visual signals)采用类中心学习来探索人脸-语音间的特征关联. 类似的LDJE[15](learning discriminative joint embeddings for efficient face and voice association)通过使用双向五元组约束、身份约束和中心约束训练网络. SSNet和LDJE这2种方法都主要通过中心约束来监督嵌入特征，不能充分地利用潜在语义学习更可靠的跨模态关联.

得益于深度学习的发展，将表示学习和聚类算法结合是深度神经网络最具前途的方法之一. 而自监督学习作为目前最热门的框架，旨在使用原始特征生成监督网络训练的伪标签，通过潜在特征关联进行学习. 深度聚类DeepCluster[16](online deep clustering for unsupervised representation learning)中引入了学习图像表示的自监督方法，通过对特征无监督聚类的结果约束图像的特征表示. 而将自监督学习运用于跨模态关联，需要考虑模态间自监督学习的可适用性以及跨模态下自监督学习生成特征的异构性.

2 跨人脸-语音自监督学习方法

本文所提出的双向伪标签自监督学习的跨人脸-

语音学习方法总体框架如图 1 所示，该框架由 2 个主要模块组成，即跨模态加权残差网络模块和自监督学习模块. 前一个模块旨在学习跨模态公共嵌入特征，生成模态间的通用特征表示；而后一个模块创新性地利用自监督学习方法生成伪标签，并将一种模态下生成的伪标签作为唯一的监督信号去监督另一种模态的特征学习，实现双向伪标签关联. 这 2 个模块相互结合进行训练，以促进人脸-语音的跨模态关联学习.

2.1 基本定义

为了方便对本文的陈述，将对变量及符号进行形式化定义. 人脸数据集和语音数据集分别用 $X^{\text{face}}=\{x_i^{\text{face}}\}_{i=1}^{N}$ 和 $X^{\text{voice}}=\{x_i^{\text{voice}}\}_{i=1}^{N}$ 表示，其中 N 表示样本总数，而 x_i^{face} 和 x_i^{voice} 表示第 i 条人脸-语音数据对. 人脸和语音对应着共享的标签集 $Y=\{y_i^c\}_{i=1}^{N}$，其中 y_i^c 表示样本 i 对应有 c 个标签类别. 通过人脸子网络和语音子网络对样本 i 的人脸-语音数据进行高级特征提取分别表示为 $D^{\text{face}}(x_i^{\text{face}})$ 和 $D^{\text{voice}}(x_i^{\text{voice}})$.

2.2 跨模态加权残差网络

人脸-语音由于模态的不同，异构特征间存在着巨大的语义鸿沟. 要想探索跨模态下人脸-语音的关联，如何跨越异构特征之间的语义鸿沟至关重要. 受多模态深度学习[17]启发，双流深度网络能兼容学习和探索异构特征间的通用表示. 现有的人脸-语音方法[9]局限于使用权值共享的单一全连接层获取通用特征，而单层的网络结构无法挖掘人脸-语音特征中的非线性相关性. 为解决这个问题，本文设计了跨模态加权残差网络模块来学习跨模态下异构特征的通用表示. 其思想是使双流深度网络和残差网络结构[18]相结合，在保留原始特征的同时，学习到人脸-语音特征间的非线性相关性. 跨模态加权残差网络结构由 2 个全连接层（fully connected layer）组成，它们的加权参数分别用 ω_1 和 ω_2 表示. 人脸或者语音数据用 $\boldsymbol{x}$ 表示，将数据经过 2 个全连接层处理表示为 $FC(\boldsymbol{x})=d(\omega_2\sigma(\omega_1\boldsymbol{x}))$，其中 $\sigma(\cdot)$ 为双曲正切激活函数 $\tanh(\cdot)$，$d(\cdot)$ 表示权重丢弃层（dropout layer），用于减少特征冗余，提高网络的泛化能力. 通过人脸和语音子网络提取的特征，将通过共享权重的相同结构进行处理，得到的人脸高级特征和语音高级特征分别定义为

$$\boldsymbol{f}_i^{*}=\sigma(D^{\text{face}}(x_i^{\text{face}})+\alpha\cdot FC(D^{\text{face}}(x_i^{\text{face}}))), \tag{1}$$

$$\boldsymbol{v}_i^{*}=\sigma(D^{\text{voice}}(x_i^{\text{voice}})+\alpha\cdot FC(D^{\text{voice}}(x_i^{\text{voice}}))), \tag{2}$$

其中 $\sigma(\cdot)$ 可用于避免训练过程中的梯度过度波动，缩放因子 α[19]是一个可学习的参数. 残差网络结构将输出特征进行跳跃连接，在缓解网络梯度消失的同时，使得原始特征得以保留. 而 2 个模态之间的全连接层

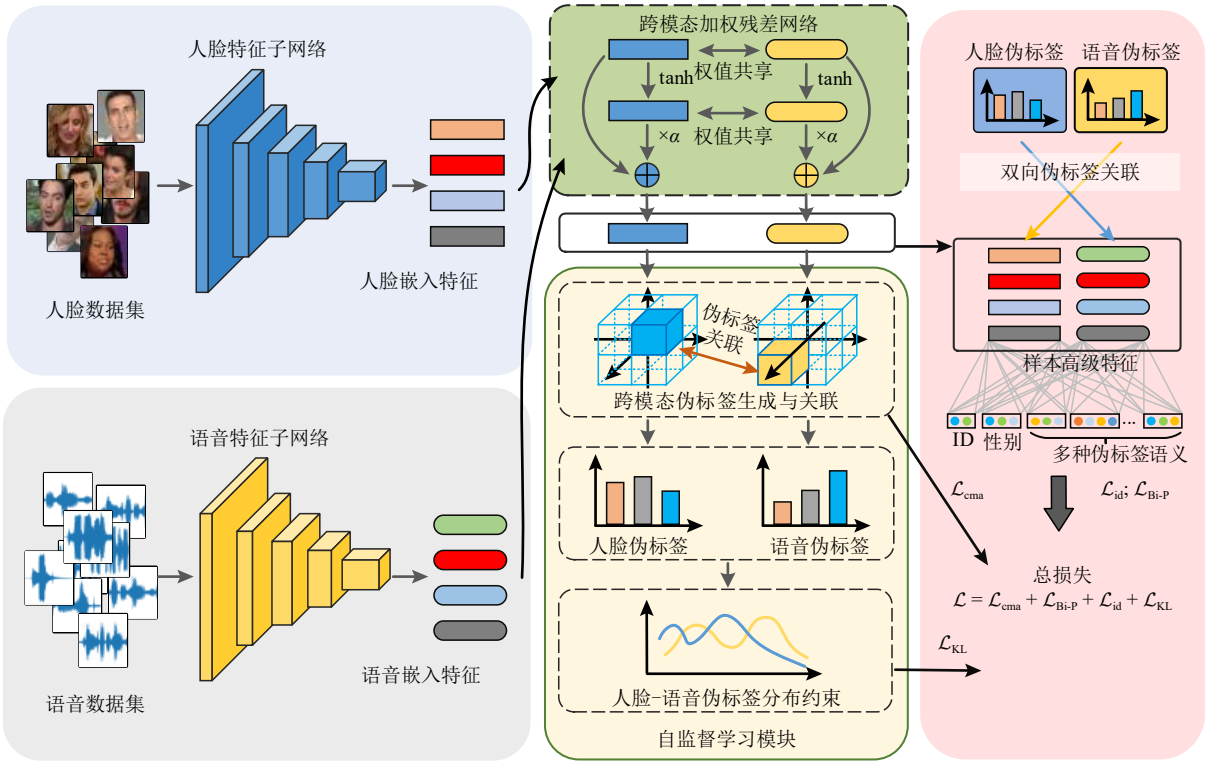


Fig. 1 The overall framework of the proposed cross-modal face-voice learning method

图 1 本文跨人脸-语音模态学习方法的总体架构

进行权值共享,有助于实现模态间的兼容性学习.基于此,通过跨模态加权残差网络,可以有效地学习人脸-语音模态间异构特征的通用表示.

2.3 基于双向伪标签关联的自监督学习

对于跨模态嵌入特征的学习,要想建立人脸-语音特征之间的关联,学习器应该将不同模态下的数据映射到一个共同的特征空间中,且需要保证来自相同身份的人脸-语音数据对在特征空间中更加接近,不同身份的人脸-语音对更加疏远.现有的大多数方法[13,15]都只是利用正则化有限的损失函数来学习人脸-语音间的跨模态对应关系,这在很大程度上忽略了人脸-语音在潜在语义上的关联.而自监督学习旨在通过探索原始的特征分布产生伪标签,进而训练模型学习潜在语义关联.因此,基于自监督学习生成的伪标签有利于捕获人脸-语音间的弱相关性.

本文所提出的架构旨在创新性地利用模态下生成的伪标签信息实现模态间双向监督,并且从每种模态中学习到的伪标签应该很好地与下游任务对应.为此,本文研究了这样一个假设,即想要捕获跨人脸-语音模态间的对应关系,可以从自监督学习下获取的伪标签中揭示出有效的潜在语义信息.而为了生成更有效的伪标签监督信号,本文还考虑了人脸-语音数据在2种模态下伪标签分布之间的对应关系.对于通过跨模态加权残差网络得到的人脸高级特征$\boldsymbol{f}_i^*$和语音高级特征$\boldsymbol{v}_i^*$,它们的特征维度都为$\mathbb{R}^{1\times K}$.因为本文采取小批量训练方案,每个小批量中包含B个样本,所以会得到一个$B\times K$维的特征矩阵$\boldsymbol{M}$.为消除矩阵$\boldsymbol{M}$中特征之间单位和尺度差异的影响,需要对$\boldsymbol{M}$进行归一化处理.定义$\bar{\boldsymbol{x}}\in\mathbb{R}^{B\times 1}$表示矩阵$\boldsymbol{M}_{(B,K)}$中的列向量,则其归一化表示为

$$\boldsymbol{x}^* = \varepsilon_1 + \frac{(\bar{\boldsymbol{x}}-\min(\bar{\boldsymbol{x}}))(\varepsilon_2-\varepsilon_1)}{\max(\bar{\boldsymbol{x}})-\min(\bar{\boldsymbol{x}})}, \tag{3}$$

其中ε_1和ε_2的取值分别为−1和1,$\max(\bar{\boldsymbol{x}})$和$\min(\bar{\boldsymbol{x}})$分别表示列向量$\bar{\boldsymbol{x}}$中的最大特征值和最小特征值.经过归一化处理后得到特征矩阵$\boldsymbol{M}^*_{(B,K)}$,基于其中B个K维特征,这里设置特征原型$\boldsymbol{\eta}\in\mathbb{R}^{1\times K}$.而$\boldsymbol{\eta}$的获取需要符合约束条件:

$$\min \sum_{\substack{\boldsymbol{x}'\text{为}\boldsymbol{M}^*\text{的列向量};\\ \boldsymbol{x}'\in\mathbb{R}^{1\times K}}} dis(\boldsymbol{\eta},\boldsymbol{x}'), \tag{4}$$

其中$dis(\cdot,\cdot)$表示欧氏空间距离.对于特征矩阵$\boldsymbol{M}^*$,为了探索特征空间中的隐式语义,通过设置聚类总数为q的无监督算法K-means,对其进行迭代聚类,直至收敛.因此,特征矩阵$\boldsymbol{M}^*$中的B个样本将被划分到q个簇中,而每个簇在迭代过程中都有其对应的中心特征$(\boldsymbol{o}_1,\boldsymbol{o}_2,\cdots,\boldsymbol{o}_q;\boldsymbol{o}_i\in\mathbb{R}^{1\times K})$.我们根据特征向量$\boldsymbol{o}_i$与特征原型$\boldsymbol{\eta}$的空间关系为簇分配伪标签,且需要保证每个簇有自己唯一的伪标签,簇与簇之间的伪标签不存在差异性,所以采用独热编码(one-hot)的方式生成伪标签,过程如图2所示.q个簇将对应大小为$\boldsymbol{L}_x\in\mathbb{R}^{1\times q}$的0,1的编码(例如:$\boldsymbol{L}_x=(0,0,1,0,0)$,$q=5$).基于簇的$q$个特征向量$\boldsymbol{o}_i$与特征原型$\boldsymbol{\eta}$的欧氏空间距离$dis(\boldsymbol{o}_i,\boldsymbol{\eta})$排序后,为距离$\boldsymbol{\eta}$最近的簇分配伪标签向量$\boldsymbol{L}_x=(1,0,0,\cdots,0)$,为距离$\boldsymbol{\eta}$最远的簇分配伪标签向量$\boldsymbol{L}_x=(0,\cdots,0,0,1)$.基于这种空间排序依次为簇分配伪标签,保证了每个簇生成的伪标签是唯一的,且伪标签之间不存在差异性.因此,批中的B个样本根据其所在的簇,通过无监督聚类及簇中心特征和特征原型空间距离约束,被分配伪标签向量$\boldsymbol{L}_x$.对于自监督学习下的人脸-语音关联,需要保持相同身份的人脸-语音数据在分配伪标签后语义的一致性,同时显示出不相关人脸-语音对的差异性.本文将从2种模态下获取的伪标签进行跨模态语义关联.假设第i个样本的人脸-语音数据,在通过伪标签分配后得到的伪标签向量分别为$\boldsymbol{L}^i_{\text{face}}$和$\boldsymbol{L}^i_{\text{voice}}$,则跨模态伪标签关联得分表示为

$$S_i = \boldsymbol{L}^i_{\text{face}}(\boldsymbol{L}^i_{\text{voice}})^{\mathrm{T}} = \begin{cases} 1, & \boldsymbol{L}^i_{\text{face}} = \boldsymbol{L}^i_{\text{voice}}, \\ 0, & \boldsymbol{L}^i_{\text{face}} \neq \boldsymbol{L}^i_{\text{voice}}. \end{cases} \tag{5}$$

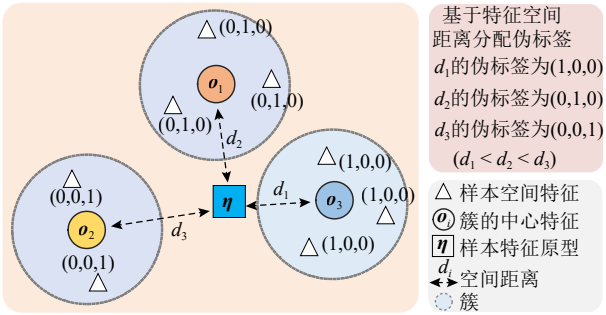


Fig. 2 Pseudo-label assignment based on feature prototype

图2 基于特征原型的伪标签分配

如当相同身份的人脸-语音样本通过伪标签分配后得到的伪标签相同时,从特征语义上说明了样本的2种模态语义更加相似,因此将给予自监督学习模块得分奖励.反之,相同样本的人脸-语音数据在2种模态下分配的伪标签不同时,学习模块则没有得分奖励.通过伪标签关联得分可以反映出自监督学习产生的人脸和语音伪标签之间的对应关系.基于此,本文构建的伪标签关联损失定义为

$$\mathcal{L}_{\text{cma}} = \frac{1}{B}\sum_{i=1}^{B}\exp(-S_i), \tag{6}$$

其中B是小批量学习的样本数，exp(·)为以e为底的指数函数. 模型在训练中，随相同身份的人脸-语音数据通过自监督学习产生的特征语义越接近，伪标签关联得分S会越高，伪标签关联损失$\mathcal{L}_{\text{cma}}$则会更小. 跨模态伪标签关联损失的减小意味着人脸-语音数据通过自监督学习生成的伪标签在跨模态下的语义一致性得以保留，同时为自监督学习的下游任务提供了稳定的伪标签监督信号.

为理解$\mathcal{L}_{\text{cma}}$损失的反向传播如何影响特征网络，本文通过获取关联得分时参数矩阵$\boldsymbol{X}$的优化进行解释，其优化过程如算法1所示.

算法1. 参数矩阵优化算法.

输入：$\boldsymbol{a}=(\phi_{\text{face}}(\boldsymbol{f}_i^*)\circ\boldsymbol{w}_{\text{face}}(\boldsymbol{L}_{\text{face}}))\in\mathbb{R}^{K\times1}$，$\boldsymbol{b}=(\phi_{\text{voice}}(\boldsymbol{v}_i^*)\circ\boldsymbol{w}_{\text{voice}}(\boldsymbol{L}_{\text{voice}}))\in\mathbb{R}^{K\times1}$，$\boldsymbol{w}_{\text{face}}$和$\boldsymbol{w}_{\text{voice}}$为伪标签向量分别对应的权重向量，$\phi_{\text{face}}$与$\phi_{\text{voice}}$分别为$\boldsymbol{f}_i^*$和$\boldsymbol{v}_i^*$的特征判别器，设$\boldsymbol{X}\in\mathbb{R}^{K\times K}$为待优化参数矩阵；

输出：$S=\boldsymbol{a}^{\text{T}}\exp(\boldsymbol{X}\boldsymbol{b})$，$\exp(\cdot)$为逐个元素求指数，方便矩阵求导，基于此对$\boldsymbol{X}$进行优化.

① $\text{d}S=\boldsymbol{a}^{\text{T}}(\exp(\boldsymbol{X}\boldsymbol{b})\odot(\text{d}\boldsymbol{X}\boldsymbol{b}))$；

②根据迹$\begin{cases}\text{tr}(\boldsymbol{A}^{\text{T}}(\boldsymbol{B}\odot\boldsymbol{C}))=\text{tr}((\boldsymbol{A}\odot\boldsymbol{B})^{\text{T}}\boldsymbol{C}),\\ \text{tr}(\boldsymbol{A}\boldsymbol{B})=\text{tr}(\boldsymbol{B}\boldsymbol{A});\end{cases}$

③根据导数微分关联$\text{d}S=\text{tr}\left(\left(\dfrac{\partial S}{\partial\boldsymbol{X}}^{\text{T}}\right)\text{d}\boldsymbol{X}\right)$；

④基于②和③对①进行优化变换，即优化

$$\frac{\partial S}{\partial\boldsymbol{X}}=(\boldsymbol{b}(\boldsymbol{a}\odot\exp(\boldsymbol{X}\boldsymbol{b}))^{\text{T}})^{\text{T}}=(\boldsymbol{a}\odot\exp(\boldsymbol{X}\boldsymbol{b}))\boldsymbol{b}^{\text{T}}.$$

其中对于$\boldsymbol{f}_i^*$和$\boldsymbol{v}_i^*$将会有判别学习器ϕ_{face}，ϕ_{voice}生成可用特征，并将其用于人脸-语音伪标签向量对应的权重向量得到$\boldsymbol{a}$，$\boldsymbol{b}\in\mathbb{R}^{K\times1}$，最后通过算法1中的迹变换和微分关联来对参数矩阵$\boldsymbol{X}$进行优化.

值得注意的是，随着当前小批量样本训练的完成，本轮的特征原型$\boldsymbol{\eta}$将保留并参与下一轮批量训练中特征原型的更新迭代：

$$\boldsymbol{\eta}^{(n)}\leftarrow\lambda\boldsymbol{\eta}^{(n)}+(1-\lambda)\boldsymbol{\eta}^{*(n-1)},\tag{7}$$

其中$\boldsymbol{\eta}^*$表示前$n-1$轮批量学习中特征原型的平均特征，参数λ=0.9. 特征原型的更新迭代保证了每轮样本的训练特征得以保留，使得更新后的特征原型更具稳定性. 且基于特征原型划分的伪标签在簇数更多时，不会受个别样本特征的干扰，从而生成的伪标签作为跨模态关联的监督信号更具鲁棒性和可解释性.

基于自监督学习方法生成的人脸-语音伪标签，本文实现跨模态伪标签信号的双向监督，进而增强人脸-语音模态间的语义关联. 对于样本i的人脸特征和语音特征，其通过自监督学习得到的人脸伪标签和语音伪标签分别为$\boldsymbol{L}_{\text{face}}^i$和$\boldsymbol{L}_{\text{voice}}^i$，将其作为对方模态下的监督信号，进而得到双向伪标签关联损失：

$$\mathcal{L}_{\text{Bi-P}}=\sum_{i=1}^{B}\{\ell(\phi_{\text{face,voice}}(\boldsymbol{f}_i^*),\boldsymbol{L}_{\text{voice}}^i)+\ell(\phi_{\text{face,voice}}(\boldsymbol{v}_i^*),\boldsymbol{L}_{\text{face}}^i)\},\tag{8}$$

其中$\ell(\cdot,\cdot)$为交叉熵损失，$\phi_{\text{face,voice}}$表示判别学习器. 双向伪标签关联损失从跨模态角度利用2个模态下的伪标签实现跨模态交叉监督，从而同时优化2个模态间的关联损失，增强相同身份的人脸-语音数据对之间的语义相关性，并扩大不相关人脸-语音对之间的差异性. 基于此，双向伪标签关联损失可以有效地挖掘模态间的潜在语义，提高跨模态下生成的人脸-语音关联表示的鲁棒性和模型的泛化能力.

2.4 辅助损失函数

为了帮助整体网络更好地学习人脸-语音之间的跨模态关联，本文设计并使用了2个辅助损失函数加速模型收敛，促进整体网络的学习.

1）身份预测损失. 据文献[5]可知，有限的监督信息能够增强人脸-语音嵌入特征的判别性，同时增强跨模态加权残差网络处理异构特征时的可分离性. 本文基于*ID*损失和性别约束，通过参数分类器来学习人脸-语音潜在语义的判别性嵌入，其中身份预测损失为：

$$\mathcal{L}_{\text{id}}=\sum_{i=1}^{B}\sum_{c\in\{ID,g\}}\{\ell(\phi_c(\boldsymbol{f}_i^*),y_i^c)+\ell(\phi_c(\boldsymbol{v}_i^*),y_i^c)\},\tag{9}$$

其中ϕ_c对应在*ID*和性别约束g下的全连接判别学习器. 该损失将用于加速网络模型的收敛，促进跨模态加权残差网络对判别性特征的学习.

2）伪标签分布损失. 为了进一步规范2种模态下来自相同身份伪标签之间的一致性，学习框架将训练中B个样本的伪标签分布视为一个整体，通过归一化函数softmax获取样本伪标签概率分布. 我们将B个样本的人脸和语音数据对应的伪标签概率分布分别表示为$p(f)$和$p(v)$，并基于KL散度（Kullback Leibler divergence）生成跨模态下的伪标签分布损失：

$$\mathcal{L}_{\text{KL}}=\alpha(F_{\text{KL}}(p(f)\|p(v))+F_{\text{KL}}(p(v)\|p(f))),\tag{10}$$

其中α=0.5，$F_{\text{KL}}(\cdot)$为KL散度计算函数. 使用$F_{\text{KL}}(p(f)\|p(v))$和$F_{\text{KL}}(p(v)\|p(f))$相结合是为了保持损失的对称性. 有且仅当自监督学习到的人脸-语音伪标签概率分布相同时，$\mathcal{L}_{\text{KL}}$=0. 最小化伪标签分布损失是从整体跨模态关联角度，通过相同身份个体的人脸-语音特征更接近，扩展到2个模态下的伪标签分布一致性，使得在自监督模块注重相同身份的人脸-语音对应性学习，从而强制深度网络学习到的跨模态关联

特征更具鲁棒性.

2.5 模型训练

本文构建的整体损失函数表示为

$$\mathcal{L}=\mathcal{L}_{\text{cma}}+\gamma_1\mathcal{L}_{\text{Bi-P}}+\gamma_2\mathcal{L}_{\text{id}}+\mathcal{L}_{\text{KL}}. \tag{11}$$

默认情况下, γ_1和γ_2的权重系数分别设置为 10 和 0.1. 跨模态伪标签关联损失$\mathcal{L}_{\text{cma}}$和伪标签分布损失$\mathcal{L}_{\text{KL}}$的权重系数都设置为 1, 一方面保证了它们在促进模态间潜在语义特征挖掘中的协同作用, 另一方面加速了整体损失函数在训练中更快地迭代与收敛. 本文将每批次训练的样本数设置为 128, 并选择结合了动量技术、RMSprop(root mean square prop)修正的 Adam[20](adaptive moment estimation)方法作为优化模型. 在训练期间, 学习率会随着训练轮数的增加而衰减, 初始的学习率设置为 10^{-3}, 衰减到的最小学习率为 10^{-8}. 值得注意的是, 在实践中, 本文通过设置不同簇数q来获取多种人脸-语音伪标签, 并在实验中通过多种伪标签组合来挖掘人脸-语音间的潜在语义关联, 从而探索出更深层的跨模态人脸-语音联系.

3 实验与结果

为了充分评估本文所提出算法的有效性, 本文在公开的 Voxceleb1[21] 和 VGGFace[22] 语音视频数据集上进行实验, 并采取基准的评价准则进行量化评估. 具体的实验细节与设置如下.

3.1 数据集

Voxceleb1 中总计包含 10 万多条音频和 2 万多条视频, 而 VGGFace 中包含 2 622 个身份信息. 在实验中, 对这 2 个数据集的数据交集共 1 225 个身份进行数据集划分, 其中训练集、验证集和测试集中包含的人物身份个数分别为 924, 112, 189. 为了保证实验评估时的有效性和鲁棒性, 本文在实验中选取的训练集和验证集以及测试集之间个体身份信息完全不相交.

3.2 实验细节

1)人脸数据处理.首先对原始检测的人脸图像进行缩放, 然后通过随机裁剪函数进行裁剪, 并统一图像大小为 224×224×3. 在训练阶段采用概率为 50% 的随机水平翻转处理. 人脸子网络使用 ResNet-34[23] 架构实现, 最终输出的人脸特征维数为 256.

2)语音数据处理. 语音数据首先通过语音检测函数清洗后除去原始音频中包含的静音片段, 然后根据语音片段时长进行裁剪. 如果语音片段时长大于 10 s, 则随机保留 10 s; 若片段时长小于 10 s, 则会随机复制增加语音长度到 10 s. 语音处理使用帧长 25 ms、帧间隔 10 ms 的梅尔倒谱系数, 并对处理后的语音片段进行归一化处理. 语音子网络采用 DIMNet-voice[5] 架构实现, 最终输出的语音特征维数为 256.

3.3 实验评价指标

为了验证本文方法的有效性, 实验将在 4 种人脸-语音跨模态匹配任务上进行测试.

1) 跨模态验证任务

跨模态验证用来判断给定的人脸数据和语音数据是否属于相同身份, 该任务使用曲线下面积(area under curve, AUC)作为唯一的评价指标.

2) 跨模态检索任务

在跨模态检索任务中将给定一种模态的待测样本, 需要从总数据集中查询与待测样本匹配的正例, 所以该任务挑战难度更大. 本任务将采用平均准确率(mean average precision, mAP)作为评价指标.

3) 1∶2 匹配任务

1∶2 匹配任务由人脸图片检索语音片段(F-V)和语音片段检索人脸图片(V-F)这 2 种情况组成. 对于 F-V 的 1∶2 匹配, 给定一张人脸图片, 需要从 2 段语音片段中判断出哪个和人脸图片身份相同. 同理可知 V-F 的 1∶2 匹配, 给定一段语音片段, 需要从 2 张人脸图片中判断出哪个和语音身份相同. 本任务中采用百分制的准确率(accuracy, ACC)作为评价指标.

4) 1∶N匹配任务

1∶N匹配任务是 1∶2 匹配任务的扩展, 其将待匹配的样本总数增加到 N, 且需要从中识别出唯一的正例. 同样地, 1∶N匹配也存在 F-V 和 V-F 的 2 种情况, 且随着样本总数 N 的增加, 任务难度也逐渐增加. 该任务也采用准确率 ACC 作为评价指标.

3.4 实验对比结果

为了验证本文所提出方法的有效性, 将通过 3.3 节中所涉及的 4 种跨人脸-语音模态匹配任务进行测试. 值得注意的是, 本文所提出的跨模态学习架构, 由于伪标签生成跟簇数q有关, 而不同的伪标签会对学习到的人脸-语音关联表示产生影响, 所以在实验中尝试了不同的伪标签组合. 本文实验中使用了簇数分别为 8, 32, 64 来生成伪标签, 其形式化标记分别对应 Bi-Pcm-F(first), Bi-Pcm-S(second), Bi-Pcm-T(third)方法. 除此之外, 本文还尝试设置了不同的伪标签组合来探索更多跨人脸-语音模态的潜在语义关联. 本文设置了 4 种伪标签组合: 1)8 和 32 组合; 2)8 和 64 组合; 3)32 和 64 组合; 4)8, 32, 64 组合. 这 4 种组合分别对应 Bi-Pcm-FS, Bi-Pcm-FT, Bi-Pcm-ST, Bi-

Pcm-FST 方法. 实验中，当不同伪标签数的方法进行组合后，伪标签的分配以及训练的过程并行执行，最后生成的整体损失也将进行叠加.

1）跨模态验证

参考文献 [5]，本文与现有方法的实验比较如表 1 所示，实验在不同分类数据上进行. 其中“U”表示人脸-语音数据对没有进行分类，“G”(gender)表示人脸-语音数据对中的 2 个测试者性别相同，“N”(nationality)表示人脸-语音数据对中的 2 个测试者的国籍相同，“A”(age)表示人脸-语音数据对中的 2 个测试者年龄相同. 而对于“GNA”这种情况，则是 2 个测试者的性别、国籍和年龄都相同. 从表 1 可知，本文所提出的 Bi-Pcm-FST 方法相比较于 PINs, SSNet 方法，实验性能在各个验证任务上平均提升 5 个百分点. 实验表明本文模型在不同的任务上都更具有效性.

2）跨模态检索

跨模态检索任务的实验结果如表 2 所示. 本文在 F-V 和 V-F 的 2 个情景上都进行了检索实验. 为了与未进行学习的特征进行对比，本文在实验中增加了随机情况下(Chance)的实验结果，Chance 方法将在跨模态检索以及 1∶N 匹配任务中使用. 方法 Bi-Pcm-FST 的平均 mAP 为 6.20，高于目前先进的 DIMNet-IG 方法将近 2 个百分点，这说明基于本文的特征表示在面对大量数据检索任务时更具健壮性.

Table 1 AUC Values of Cross-Modal Verification Task

表 1 跨模态验证任务的 AUC 值

方法	U	G	N	A	GNA
PINs[10]	78.5	61.1	77.2	74.9	58.8
SSNet[3]	78.8	62.4	53.1	73.5	51.4
DIMNet-I[5]	82.5	71.0	81.9	77.7	62.8
DIMNet-IG[5]	83.2	71.2	81.9	78.0	62.8
本文（Bi-Pcm-FST）	**85.0**	**71.2**	**84.3**	**79.6**	**64.7**

注：U 为未分类，G 以性别分类，N 以国籍分类，A 以年龄分类，GNA 以性别、国籍和年龄共同分类. 黑体数值表示最佳结果.

Table 2 Performance mAP of Cross-Modal Retrieval

表 2 跨模态检索中 mAP 的性能

方法	Chance	F-V	V-F	平均值
FV-CME[11]	0.46	2.18	1.96	2.07
VFMR3[24]	2.15		5.00	
DIMNet-I[5]	1.07	4.17	4.25	4.21
DIMNet-IG[5]	1.07	4.23	4.42	4.33
本文（Bi-Pcm-FST）	1.01	**6.04**	**6.36**	**6.20**

注：F-V 为人脸图片检索语音片段，V-F 为语音片段检索人脸图片，平均表示 F-V 和 V-F 的平均值. 黑体数值表示最佳结果.

3）1∶2 匹配

1∶2 匹配在不同分类数据上的测试结果如表 3 所示，其中数据分组“U”“G”“N”的方式同本节跨模态检索中的描述一致. 此任务共包括 2 种情景，分别为 F-V 和 V-F. 本文基于不同伪标签组合的 Bi-Pcm 方法，在 2 种情景下进行了多组实验以探索多种伪标

Table 3 ACC on Cross-Modal 1∶2 Matching Task

表 3 跨模态 1∶2 匹配任务的准确率 %

方法	F-V				V-F			
	U	G	N	GN	U	G	N	GN
SVHF[9]	79.50	63.40			81.00	63.90		
FV-CME[11]	77.80	60.80			78.10	61.70		
LAFV[12]	78.60	61.60			78.20	62.90		
PINs[10]	83.80							
DIMNet-I[5]	83.52	71.78	82.41	70.90	83.45	70.91	81.87	69.89
DIMNet-IG[5]	84.03	71.65	82.96	70.78	84.12	71.32	82.65	70.39
LDJE[15]	85.42	**73.52**	84.48	71.11	85.18	**74.29**	83.97	70.70
Bi-Pcm-F（本文）	84.81	71.93	83.81	70.89	84.77	72.08	83.56	70.53
Bi-Pcm-S（本文）	84.65	72.05	83.96	71.07	84.80	72.11	83.72	70.77
Bi-Pcm-T（本文）	85.13	72.22	84.07	71.12	84.82	72.37	83.86	70.69
Bi-Pcm-FS（本文）	85.27	72.28	84.25	71.08	85.11	72.55	84.02	70.78
Bi-Pcm-FT（本文）	85.34	72.46	84.44	71.14	85.23	72.94	84.17	70.84
Bi-Pcm-FST（本文）	**85.83**	73.01	**85.00**	**71.45**	**85.69**	73.33	**84.26**	**71.10**

注：F-V 为人脸图片匹配语音片段，V-F 为语音片段匹配人脸图片，U 表示未分类，G 表示以性别分类，N 表示以国籍分类，GN 表示以性别和国籍分类. 黑体数值表示当前任务中的最佳结果.

签语义对人脸-语音关联的影响. 从实验结果可知，本文基于 Bi-Pcm-FST 的多伪标签组合相比较其他伪标签组合在多种 1∶2 匹配任务上性能表现更佳，所以本文中其他的对比实验均以 Bi-Pcm-FST 作为代表. 在 1∶2 匹配任务中，Bi-Pcm-FST 与目前主流的 LDJE 相比虽然只获得了少量的提升，但是 LDJE 方法在训练中使用了大量的人为监督标签来构造双向五元组约束，并利用中心约束以及身份约束，本质上过度依赖有监督学习，况且监督标签的获取成本昂贵且十分耗时. 而 Bi-Pcm-FST 更注重自监督学习生成可用伪标签来代替这些传统的有监督标签，且取得了更好的性能表现，这种获取可用伪标签的方法为跨人脸-语音模态的研究开创了一种更加新颖的思维. 跨模态 1∶2 匹配的实验结果也表明，本文基于双向伪标签关联的自监督学习能够为人脸-语音探索出更多的潜在语义信息.

4) 1∶N 匹配

1∶N 匹配结果如图 3 所示. 此项任务随待匹配样本数 N 的增加，实验难度也进一步增大. 可以发现各项工作的准确率也随 N 的增加而逐渐降低. 但是 Bi-Pcm-FST 方法在 V-F 和 F-V 两种情景下，与其他主流方法相比，依然具有更好的表现. 由准确率曲线可以发现，Bi-Pcm-FST 方法随待匹配样本数 N 的增加，匹配准确率相比较其他方法衰减得更加平缓，即使在 V-F 的 1∶N 匹配任务中难度较大的“G”分组上，当 N=6 时，匹配准确率也能比主流的 DIMNet 方法提高 2 个百分点. 通过 1∶N 匹配任务的实验结果进一步说明本文架构具有更强的潜在语义挖掘能力.

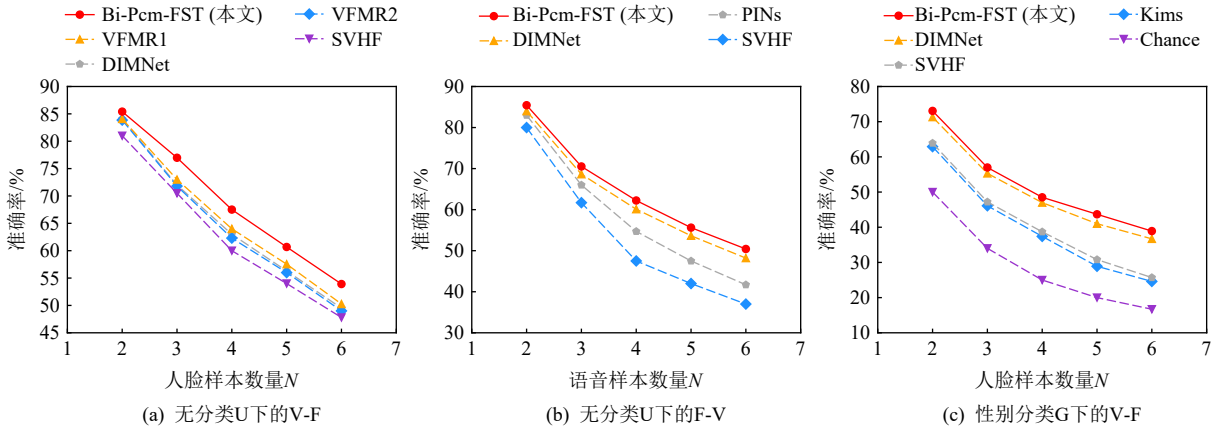


Fig. 3 Comparison of cross-modal 1∶N matching performance

图 3 跨模态 1∶N 匹配的性能对比

4 消融实验与分析

4.1 不同学习模块的消融实验

本文所提出的框架主要由 2 个模块组成，即跨模态加权残差网络模块和基于双向伪标签关联的自监督学习模块. 使用不同模块的消融实验如表 4 所示，其中 CMWR 表示跨模态加权残差网络，id 表示可用语义信息的嵌入，self-learn 表示自监督学习模块. 从表 4 中各个模块的消融实验可以发现，当单独使用跨模态加权残差网络或者单独使用自监督学习模块时，虽然整体网络的性能都能有所提升，但是提升幅度很小，例如在跨模态验证任务的“U”分组上只能提升 1.7 个百分点. 但将 2 种模块进行结合后，整体性能在“U”分组上提升 4 个百分点，说明 2 个模块之间的相互协作对促进整体网络的性能表现有重要的影响. 依次来看，跨模态加权残差网络能够跨越模态间语义鸿沟，从而有效地学习人脸-语音间的关联表示；而基于双向伪标签关联的自监督学习模块可以生成高效伪标签来促进整体网络性能的提升.

Table 4 Ablation Studies of Cross-Modal Verification

表 4 跨模态验证上的消融实验

方法	U	G	N	A	GNA
id	81.2	67.4	80.6	77.5	61.1
id+self-learn	82.7	68.8	82.0	78.6	62.1
CMWR+id	82.9	69.5	82.7	78.4	63.3
CMWR+id+self-learn	**85.0**	**71.2**	**84.3**	**79.6**	**64.7**

注：U 为未分类，G 以性别分类，N 以国籍分类，A 以年龄分类，GNA 以性别、国籍和年龄共同分类. 黑体数值表示当前任务中的最佳结果.

4.2 各项损失函数的消融实验

在本文中，损失函数是用来约束人脸-语音特征表示的关键因素. 因此，实验中进一步研究了损失函数对跨模态匹配性能的影响，图 4 展示了不同损失

函数对 F-V 跨模态 1∶2 匹配任务的消融结果. 需要注意, 双向伪标签关联损失 $\mathcal{L}_{\text{Bi-P}}$和伪标签分布损失 $\mathcal{L}_{\text{KL}}$的构成都需要跨模态伪标签关联损失 $\mathcal{L}_{\text{cma}}$的协助, 所以无法进行将 $\mathcal{L}_{\text{cma}}$单独移除的实验. 从消融结果可以发现, 当总体网络缺少 $\mathcal{L}_{\text{Bi-P}}$时, 实验准确率下降得最为明显, 总体性能下降 1.4 个百分点, 说明双向伪标签关联约束对促进网络性能提升有着重要作用. 消融实验中, 移除跨模态分布损失 $\mathcal{L}_{\text{KL}}$后, 整体网络性能轻微下降了 0.4 个百分点. 而 $\mathcal{L}_{\text{cma}}$通过得分奖励机制, 使得自监督学习模块生成人脸-语音伪标签, 进而参与 $\mathcal{L}_{\text{Bi-P}}$和 $\mathcal{L}_{\text{KL}}$来约束跨模态特征学习. 因此, 在移除 $\mathcal{L}_{\text{cma}}$后将无法得到伪标签. 除此之外, 从图 4 中关于 $\mathcal{L}_{\text{cma}}$的单独消融实验可知, 当使用 $\mathcal{L}_{\text{cma}}$时整体网络性能只有微小的提高, 其原因是只基于 $\mathcal{L}_{\text{cma}}$产生的伪标签并没有被用于下游任务中, 而将 $\mathcal{L}_{\text{cma}}$生成伪标签用于 $\mathcal{L}_{\text{Bi-P}}$或 $\mathcal{L}_{\text{KL}}$时, 整体网络性能才能有不错的提升, 说明 $\mathcal{L}_{\text{cma}}$更多的作用是辅助获取高效稳定的跨模态伪标签用于下游任务的学习.

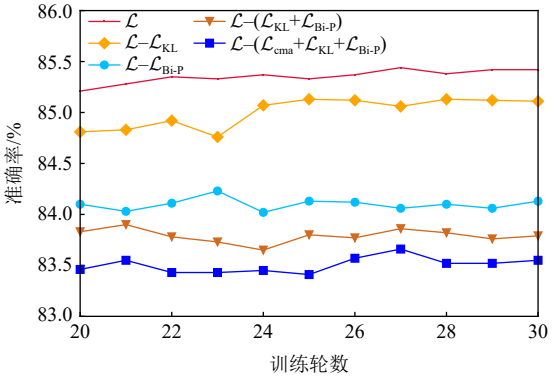


Fig. 4 Ablation studies of loss function on 1∶2 matching task

图 4 在 1∶2 匹配任务上损失函数的消融实验

综上所述, $\mathcal{L}_{\text{cma}}$帮助自监督模块生成高效伪标签用于下游任务, $\mathcal{L}_{\text{Bi-P}}$将利用上游伪标签挖掘潜在语义关联, 而 $\mathcal{L}_{\text{KL}}$将辅助 $\mathcal{L}_{\text{Bi-P}}$提高特征关联的有效性. 跨模态匹配任务的实验表现和消融结果说明了本文的多种损失相互协助, 相比较现有的方法, 可在多种跨人脸-语音匹配任务上取得更佳的性能表现.

4.3 可视化实验结果

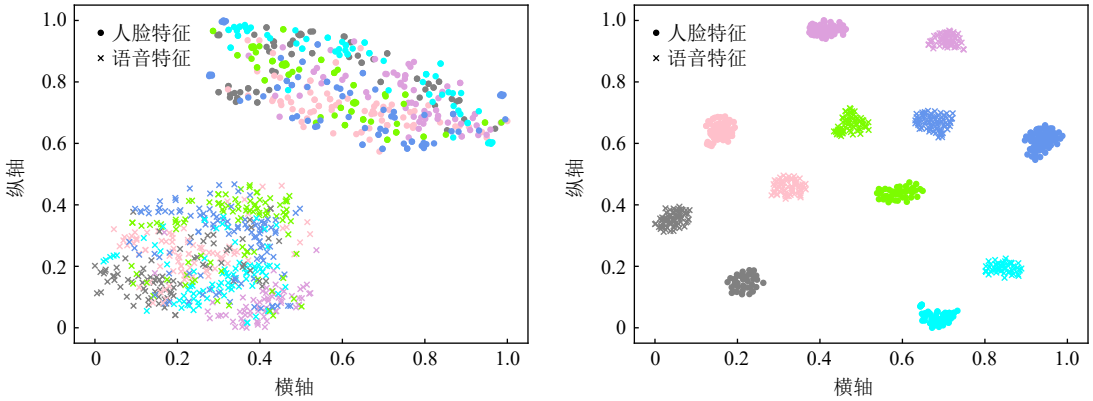
对于跨模态检索任务, 具有代表性的 V-F 检索结果如图 5 所示, 其中与语音身份相同的人脸图片已由加粗方框标注. 从跨模态检索结果可以发现, 即使待检索样本规模为整个数据集时, 本文在跨模态检索任务上依然取得了不错的性能表现.

Fig. 5 Cross-modal retrieval results

图 5 跨模态检索结果

此外, 如图 6 所示, 本文进一步利用 t-SNE[25](t-distributed stochastic neighbor embedding)算法对学习到的高维人脸-语音嵌入特征进行 2 维可视化, 其中相同颜色的数据点来自同一身份样本. 图 6(a)是初始特征分布, 人脸特征与语音特征由于模态间差异, 被划分为 2 类, 但是模态内的这 2 种特征却因没有进行辨别性学习而被混淆在一起. 图 6(b)是通过本文方法学习后的结果, 可以明显看出相同身份的人脸和语音特征的空间分布更为接近, 且不同身份的特征之间更加地疏远, 说明本文中基于双向伪标签关联的自监督学习方法确实能学习到更具判别性的跨模态特征.

为了验证 $\mathcal{L}_{\text{cma}}$损失可实现跨模态数据的编码, 本文进行了人脸-语音伪标签相似度匹配实验. 如图 7



(a) 初始特征分布 (b) 学习后的特征分布

Fig. 6 Visualization of embedding characteristics on t-SNE

图 6 嵌入特征的 t-SNE 可视化

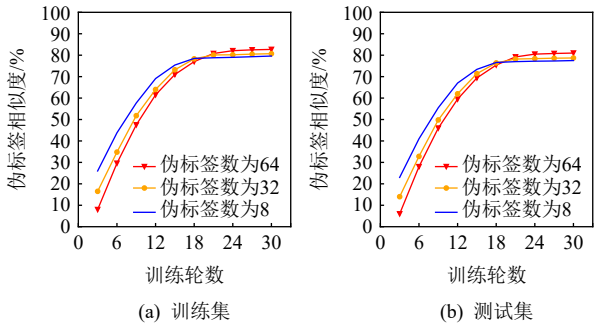


Fig. 7 Face-voice pseudo-label similarity based on cma-loss

图 7 基于$\mathcal{L}_{cma}$损失的人脸-语音伪标签相似度

所示，我们在训练集和测试集上分别对样本的人脸-语音伪标签进行了相似度统计. 本文共用到 3 种伪标签数：8, 32, 64. 伪标签数为 8 时，虽然实验收敛得更快，但是最后得到的伪标签相似度低(准确率约 77%)；伪标签数为 64 时，网络虽然收敛更慢，但是获得的伪标签相似度更高(准确率约 81%). 综上，$\mathcal{L}_{cma}$损失可以帮助实现高效的跨模态数据编码.

为了验证本文$\mathcal{L}_{Bi\text{-}P}$对模型泛化能力的影响，本文在现有的训练集基础上减少了 100 个人脸-语音数据进行模型重新训练，并在测试集上评估. 泛化能力评估实验结果如图 8 所示，当使用全部的损失后，V-F 的 1∶2 匹配任务上实验准确率只下降了 0.6 个百分点，但是在移除$\mathcal{L}_{Bi\text{-}P}$损失后，实验准确率下降了 1.3 个百分点，说明$\mathcal{L}_{Bi\text{-}P}$能保证模型的泛化能力尽可能得到保留，验证了$\mathcal{L}_{Bi\text{-}P}$能够提高跨模态下生成的人脸-语音关联表示的鲁棒性和模型的泛化能力.

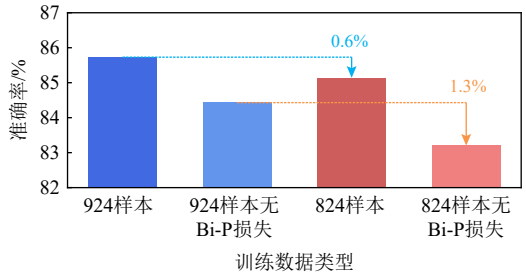


Fig. 8 Generalization ability assessment on 1∶2 matching task

图 8 在 1∶2 匹配任务上的泛化能力评估

5 总 结

本文提出了基于双向伪标签自监督学习的方法，该方法可有效地用于人脸-语音跨模态关联和匹配. 首先构建了跨模态加权残差网络来学习人脸-语音间的共享嵌入，然后创新性地提出双向伪标签关联方法生成高效伪标签，并用其监督人脸-语音实现潜在语义学习. 本文获得的模态间增强语义嵌入可适用于各种人脸-语音匹配任务. 与现有工作比较，本文在大量跨模态匹配任务中都取得了最佳的性能表现.

作者贡献声明：朱明航负责算法设计与实验；柳欣负责模型优化和算法分析；于镇宁负责模型可行性分析；徐行负责算法优化；郑书凯负责实验多样性分析.

参 考 文 献

[1] McGurk H, MacDonald J. Hearing lips and seeing voices[J]. Nature, 1976, 264(5588): 746−748

[2] Ellis A W. Neuro-cognitive processing of faces and voices[M]// Handbook of Research on Face Processing. Amsterdam : Elsevier, 1989: 207−215

[3] Nawaz S, Janjua M K, Gallo I, et al. Deep latent space learning for cross-modal mapping of audio and visual signals[C/OL] //Proc of the 14th Int Conf on Digital Image Computing: Techniques and Applications. Piscataway, NJ: IEEE, 2019[2020-05-12]. https://ieeexplore.ieee.org/document/8945863

[4] Liu Xin, Geng Jiajia, Ling Haibin, et al. Attention guided deep audio-face fusion for efficient speaker naming[J]. Pattern Recognition, 2019, 52(88): 557−568

[5] Wen Yandong, Ismail M A, Liu Weiyang, et al. Disjoint mapping network for cross-modal matching of voices and faces[J]. arXiv preprint, arXiv: 1807.04836, 2018

[6] Zhang Lu, Wang Huabin, Tao Liang, et al. Cross-media clustering by share and private information maximization[J]. Journal of Computer Research and Development, 2018, 55(1): 151−162 (in Chinese)
(张露, 王华彬, 陶亮, 等. 基于分类距离分数的自适应多模态生物特征融合[J]. 计算机研究与发展, 2018, 55(1): 151−162)

[7] Hermans A, Beyer L, Leibe B. In defense of the triplet loss for person re-identification[J]. arXiv preprint, arXiv: 1703.07737, 2017

[8] Wells T, Baguley T, Sergeant M, et al. Perceptions of human attractiveness comprising face and voice cues[J]. Archives of Sexual Behavior, 2013, 42(5): 805−811

[9] Nagrani A, Albanie S, Zisserman A. Seeing voices and hearing faces: Cross-modal biometric matching[C]// Proc of the 31st IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2018: 8427−8436

[10] Nagrani A, Albanie S, Zisserman A. Learnable PINs: Cross-modal embeddings for person identity[C]// Proc of the 15th European Conf on Computer Vision. Berlin: Springer, 2018: 71−88

[11] Horiguchi S, Kanda N, Nagamatsu K. Face-voice matching using cross-modal embeddings[C]// Proc of the 26th ACM Int Conf on Multimedia. New York: ACM, 2018: 1011−1019

[12] Kim C, Shin H V, Oh T H, et al. On learning associations of faces and voices[C]//Proc of the 14th Asian Conf on Computer Vision. Berlin: Springer, 2018: 276−292

[13] Wang Kaiye, He Ran, Wang Wei, et al. Learning coupled feature

spaces for cross-modal matching[C]//Proc of the IEEE Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2013: 2088−2095

[14] Yan Xiaoqiang, Ye Yangdong. Cross-media clustering by share and private information maximization[J]. Journal of Computer Research and Development, 2019, 56(7): 1370−1382 (in Chinese)
(闫小强, 叶阳东. 共享和私有信息最大化的跨媒体聚类[J]. 计算机研究与发展, 2019, 56(7): 1370−1382)

[15] Wang Rui, Liu Xin, Cheung Yiuming, et al. Learning discriminative joint embeddings for efficient face and voice association[C]//Proc of the 43rd Int ACM SIGIR Conf on Research and Development in Information Retrieval. New York: ACM, 2020: 1881−1884

[16] Zhan Xiaohang, Xie Jiahao, Liu Ziwei, et al. Online deep clustering for unsupervised representation learning[C]//Proc of the 30th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 6688−6697

[17] Zhen Liangli, Hu Peng, Wang Xu, et al. Deep supervised cross-modal retrieval[C]// Proc of the 32nd IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2019: 10394−10403

[18] He Kaiming, Zhang Xiangyu, Ren Shaoqi, et al. Deep residual learning for image recognition[C]// Proc of the 29th IEEE Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2016: 770−778

[19] Chung J S, Zisserman A. Out of time: Automated lip sync in the wild[C]// Proc of the 9th Asian Conf on Computer Vision. Berlin: Springer, 2016: 251−263

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

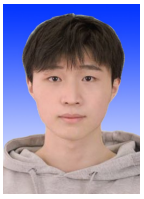
[21] Nagrani A, Chung J S, Zisserman A. Voxceleb: A large-scale speaker identification dataset[J]. arXiv preprint, arXiv: 1706.08612, 2017

[22] Parkhi O M, Vedaldi A, Zisserman A, et al. Deep face recognition[C]// Proc of the 26th Brithsh Machine Vision Conf. Durham, UK: BMVA, 2015: 1−12

[23] Venkatesh G, Nurvitadhi E, Marr D. Accelerating deep convolutional networks using low-precision and sparsity[C]//Proc of the 42nd Int IEEE Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2017: 2861−2865

[24] Xiong Chuyuan, Zhang Deyuan, Liu Tao, et al. Voice-face cross-modal matching and retrieval: A benchmark [J]. arXiv preprint, arXiv: 1911.09338, 2019

[25] Maaten L, Hinton G. Visualizing data using t-SNE[J]. Journal of Machine Learning Research, 2008, 86(9): 2579−2605



Zhu Minghang, born in 1999. Master candidate. Student member of CCF. His main research interests include multimedia analysis, pattern recognition, and deep learning.
朱明航, 1999 年生. 硕士研究生. CCF 学生会员. 主要研究方向为多媒体分析、模式识别和深度学习.

Liu Xin, born in 1982. PhD, professor. Senior member of CCF. His main research interests include information retrieval, pattern recognition, and deep learning.
柳　欣, 1982 年生. 博士, 教授. CCF 高级会员. 主要研究方向为信息检索、模式识别、深度学习.

Yu Zhenning, born in 1998. Master candidate. His main research interests include multimedia analysis, pattern recognition, and deep learning.
于镇宁, 1998 年生. 硕士研究生. 主要研究方向为多媒体分析、模式识别、深度学习.

Xu Xing, born in 1988. PhD, professor. His main research interests include multimedia information processing and security, cross-media analysis, and computer vision.
徐　行, 1988 年生. 博士, 教授. 主要研究方向为多媒体信息处理与安全、跨媒体分析、计算机视觉.

Zheng Shukai, born in 1992. Master, engineer. His main research interests multimedia data analysis and pattern recognition.
郑书凯, 1992 年生. 硕士, 工程师. 主要研究方向为多媒体数据分析、模式识别.