真实场景下使用多模态数据进行说话人识别

Saqlain Hussain Shah[1], Muhammad Saad Saeed[2],Shah Nawaz[3],Muhammad Haroon Yousaf[1]{saqlain.hussain, saad.saeed, haroon.yousaf}@uettaxila.edu.pk, shah.nawaz@desy.de[1]塔克西拉工程与技术大学， [2]NCRA群体机器人实验室，[3]德国电子同步加速器DESY

arXiv:2302.13033v1 [cs.SD] 25 Feb 2023

摘要—**近年来，借助来自YouTube的大规模视听信息，人们建立了名人与其面部和声音之间的关联。大规模视听数据集的可用性，为开发基于标准卷积神经网络的说话人识别方法提供了重要支持。因此，本文旨在利用大规模视听信息来提升说话人识别任务的性能。为实现这一目标，我们提出了一种双分支网络，用于在多模态系统中学习人脸与声音的联合表征。随后，从双分支网络中提取特征，以训练用于说话人识别的分类器。我们在名为VoxCeleb1的大规模视听数据集上评估了所提框架。结果表明，加入面部信息提升了说话人识别的性能。此外，结果还表明，面部与声音之间存在重叠。**

索引词—**说话人身份识别、多模态、人脸-语音关联**

I. 引言

说话人识别是语音处理的一项基础任务，应用于各种现实世界领域。然而，由于内在变化和外在变化，在现实世界场景下，说话人识别任务极具挑战性。内在变化与说话人属性有关，即性别、年龄和说话方式；而外在变化则包括说话人个性之外的因素，例如背景噪声、麦克风噪声等。 [1]。这使得语音信号容易产生很大程度的变化。近年来，卷积神经网络（CNN）为说话人识别任务开辟了新的路径：将语音信号转换为频谱图，再利用这些网络进行分类 [2], [3]。尽管如此，基于卷积神经网络的说话人识别方法已经超越了传统方法 [2]。然而，这些方法在现实世界场景下会出现性能下降。最近，人们构建了大规模数据集VoxCeleb1和VoxCeleb2，用于说话人识别任务。这些数据集对于开发基于卷积神经网络的说话人识别方法发挥了重要作用。例如，文献 [2], [3]修改了标准卷积神经网络，如VGG-M [4] 和ResNet [5]，以执行说话人识别任务。此外，VoxCeleb1和VoxCeleb2数据集都包含视觉信息，这对于开发各种多模态应用十分重要，例如面部与声音之间的跨模态迁移 [6]–[10], 、情感识别 [11], 、语音分离 [12]以及人脸生成 [13]。这些应用对于建立说话人人脸与声音之间的关联具有重要作用。此外，一个经过充分研究的事实是，由于声音和人脸的神经认知通路具有相同的结构 [14]，人类最终会将人们的声音与面孔联系起来。由于大规模音频-视觉数据集（如VoxCeleb1）的可用性，以及说话人人脸与声音之间的关联，一个基本问题随之产生：能否利用音频-视觉信息来改进说话人识别任务？为研究这一问题，我们提出了一个双分支网络，用于建立人脸与声音之间的关联。所提双分支网络由以下三个部分组成：1）利用面向任务的预训练子网络提取人脸和声音的特征；2）通过一系列面向人脸和声音的全连接层，学习联合多模态表征；3）损失函数形式。随后，我们提取音频片段的特征，训练用于说话人识别任务的分类器。结果表明，面部信息结合语音片段有助于改进说话人识别任务。图1展示了所提框架的训练与测试策略。

我们的主要贡献总结如下：1）我们提出了一种双分支网络，用于学习说话人脸部与声音的多模态判别联合表示。2）我们对比了仅使用语音片段与使用多模态信息的说话人识别任务。3）我们的结果表明，多模态信息显著提升了说话人识别任务的效果。

本文结构安排如下：第二节详细概述相关工作。第三节介绍所提框架，并在第四节给出结果与讨论。最后，第五节总结本文工作。

II. 相关工作

我们总结了与说话人识别和人脸-语音关联任务相关的先前研究。

A. 说话人识别

Sandra 等人 [15]为说话人识别系统奠定了基础。这类系统通过使用滤波器组和数字频谱图，尝试寻找两段语音信号之间的相似度度量。我们简要概述了说话人识别方法，并将其分为两大类：传统方法和

979-8-3503-2212-5/23/$31.00 ©2023 IEEE

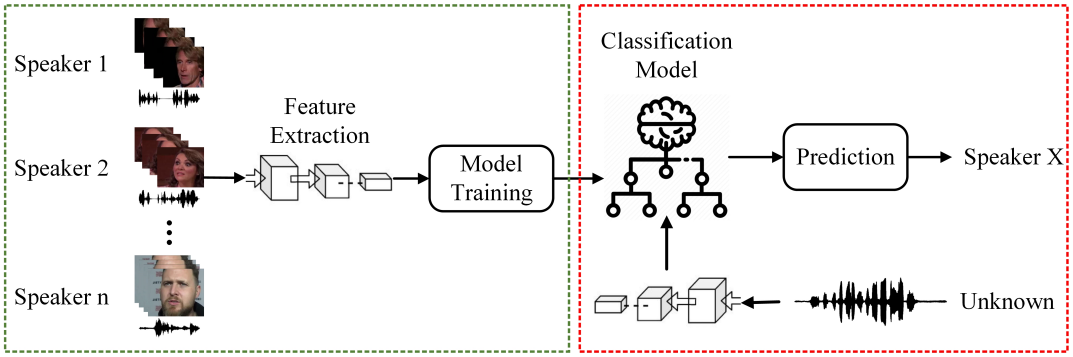


图1. 所提出研究的训练与测试策略。（绿色）显示用于训练模型的人脸轨迹。训练阶段同时使用音频和视觉模态；（红色）该阶段仅提供音频模态。该方案有助于了解一种模态对另一种模态性能的影响。

基于神经网络的方法。

传统方法。随着数据和计算资源的日益丰富，说话人识别任务取得了许多进展。然而，噪声环境仍然构成了严峻挑战。多年来，标准的说话人识别任务依赖于基于特征的方法，而这些特征需要人工干预和领域知识。这包括从语音信号的低维短时表示中提取的特征，例如梅尔频率倒谱系数 [16]。这些系统在真实环境中的性能会下降 [17], [18]。这类系统依赖人类提取有用特征的能力，这构成了系统的局限性。联合因子分析通过将语音信号分解为一组潜在因子，捕捉语音信号中说话人特定和会话特定的变异性 [19]。支持向量机（SVM）分类器在鲁棒识别任务中表现出色。然而，这类方法速度很慢、复杂度高，而且应用于各种真实场景时容易性能下降。尽管取得了这些进展，传统方法在存在噪声时的性能仍会下降。此外，随着数据规模增大，其性能也会下降。在实际应用中，环境噪声、所用传输信道以及背景中的说话人数往往都未知。在这种情况下，传统方法的性能可能会下降。

深度学习方法。过去几年，计算资源和神经网络的发展带来了更高效的方法。随着这些进步，卷积神经网络被广泛用于说话人识别等任务。例如，[2], [3] 中的工作提出了一种基于卷积神经网络的方法，将语音片段转换为频谱图，以完成说话人识别任务。随着这一进展，说话人识别任务已从手工提取特征转向数据驱动方法。具体而言，[2] 中的工作在直接从语音片段提取的频谱图上训练了一个改进的 VGG-M。

B. 人脸-语音关联

近期，研究者通过利用跨模态验证和匹配任务 [6], [7], [9], [10], [20], [21] 建立了说话者人脸与声音之间的关联。文献 [7] 使用三元组网络，为人脸—声音关联任务学习联合表示。类似地，文献 [22] 使用三元组网络 [23]，通过从人脸子网络 [24] 和语音子网络 [25] 中提取特征，最小化人脸与声音之间的距离。Nawaz 等人 [6] 利用单流网络中的类别中心学习共享潜在空间，从而无需使用成对样本或三元组样本。基于类似思路，Saeed 等人 [20], [21] 提出了一种轻量级的即插即用机制，利用人脸和声音中的互补线索构建丰富的融合嵌入，并通过正交约束，根据身份标签对其进行聚类。

与现有方法不同，我们的目标是从一个在用于说话人识别任务的人脸和声音上训练的多模态系统中提取鲁棒特征。

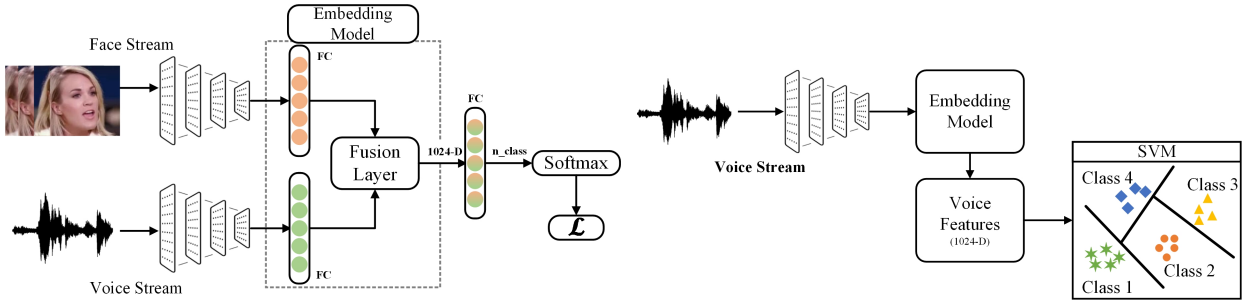
III. 总体框架

A. 基线

我们使用 VGGVox 子网络提取 VoxCeleb1 数据集的 1024 维特征，以建立基线。我们使用这些特征训练支持向量机分类器，以完成说话人识别任务。对于使用支持向量机进行多分类，决策函数的形式设为一对一，核参数设为多项式核，且多项式核函数的次数设为 3。使用 VGGVox 提取的特征训练支持向量机后，身份识别准确率为 91%。

B. 双分支网络

我们所提出的方法包括：使用双分支网络，利用面部与声音信息训练多模态系统。随后，使用该多模态系统提取特征，以训练用于说话人识别任务的分类器。面部特征和音频特征分别从 VGGFace [26] 和 VGGVox [2] 子网络中提取。随后，面部与语音特征分别输入两个独立分支，每个分支对应一种特定模态。两个子网络提取的特征经过全连接层和归一化层后进行融合。图 2 展示了所提框架。



(a) 提出的双分支网络　　(b) 单模态测试策略

图2。(a) 利用独立的模态专属嵌入网络进行即插即用的特征提取（方框）。所提出的**双分支模型**具有独立的模态专属全连接层。使用**逐元素乘法**融合两个分支。(b) 测试阶段仅使用音频数据。视觉数据设为 0。提取训练集和测试集划分中音频样本的特征。随后，使用这些特征训练支持向量机，以报告% 准确率。

C. 多模态融合

我们从面部与语音信息中提取特征。随后将这些特征融合，并传入全连接层，以学习面部和声音信号的联合表征。融合后，使用 Softmax层学习输出类别。Softmax函数用作激活函数，以预测多项概率分布；在多分类问题中，需要获得相应的概率。随后，使用从该双分支网络中提取的特征训练分类器。

D. 损失函数构造

我们希望融合特征能够捕捉说话人或身份的语义。换言之，这些特征应能够以较高的准确率预测身份标签。若属于同一类别的样本彼此靠近，而属于不同类别的样本彼此远离，则有望实现这一目标。实现这一目标的一种常用方法是采用 softmax 交叉熵（CE）损失，它也能实现稳定且高效的训练。现在，利用融合嵌入计算损失如下：

$$\mathcal{L}_{CE} = \sum_{i}^{C} \mathbf{l}_i log(f(\mathbf{l}_i), \tag{1}$$

分类交叉熵是衡量两个离散概率分布彼此可区分程度的非常好的一种指标 [27]。采用 Adam 作为优化器，学习率范围为 0.01 至 0.13。网络训练所使用的批量大小为 512、1024、2048 和 4096。当学习率为 0.04、批量大小为 2048 时，取得了最佳结果。

IV. 实验与结果讨论

A. 训练细节与数据集

数据集。VoxCeleb1 是一个大规模的视听人类语音视频数据集，这些视频从 YouTube 上以“自然场景”方式提取。视频中包含真实环境噪声，包括背景交谈声、重叠语音、笑声和录音设备噪声。表I展示了该数据集的统计信息。

表I. VoxCeleb1身份识别划分

	训练	Test	总计
说话人数	1,251	1,251	1,251
视频数	21,245	1,251	22,496
话语数量	145,265	8,251	153,516

训练。受 [20]，启发，我们提出了一种双分支网络，用于分析多模态信息对说话人识别任务的影响。人脸嵌入提取自预训练的 VGGFace [26]，而音频嵌入则提取自 VGGVox [2]。将面部与声音嵌入作为输入传入双分支网络，其子网络包含多个全连接层，后接随机失活层和归一化层。训练过程中采用了 10% 和 20% 的随机失活率。随后，将两个子网络输出的归一化嵌入进行融合，并传入全连接层和归一化层，再送入 Softmax层；该层包含 1251 个隐藏单元，对应数据集中的类别。训练时采用了多种间隔值、随机失活率、批量大小、损失函数和学习率。

测试集 从双分支网络中提取特征，用于训练和测试支持向量机分类器。我们从模型的融合层中提取了维度为 1024 的特征。特征提取分为两种方式：

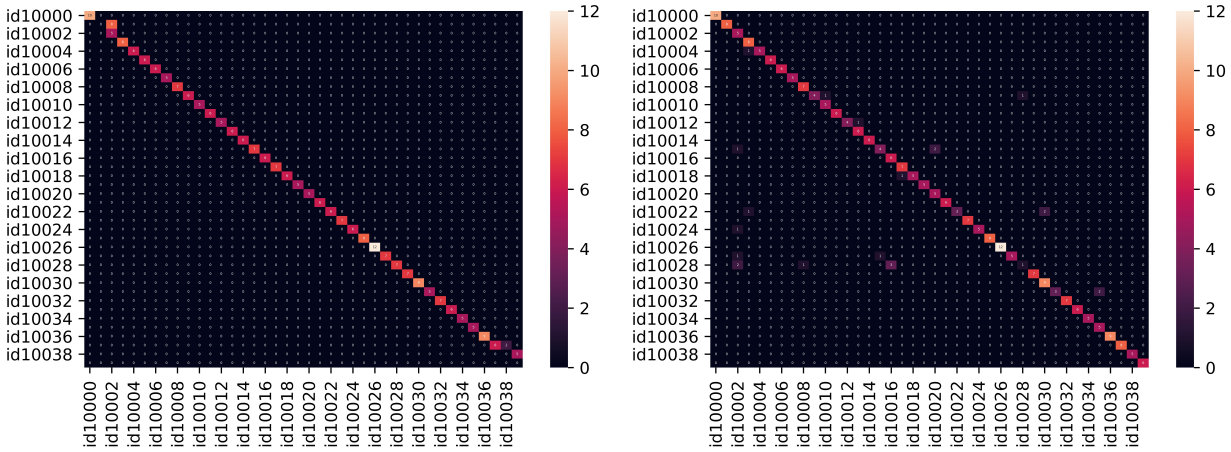
- 借助人脸信号： 在此阶段，将人脸和语音信号输入训练好的双分支网络，并从中提取语音特征。

- 不借助人脸信号： 在此阶段，仅将语音信号输入训练好的双分支网络，并从中提取语音特征。人脸子网络的输入向量设为零。

说话人身份识别。 对提取的特征进行归一化，并用于训练和测试支持向量机分类器。支持向量机的核参数设为多项式，决策函数形状设为一对一。训练时，其余参数均采用默认值。

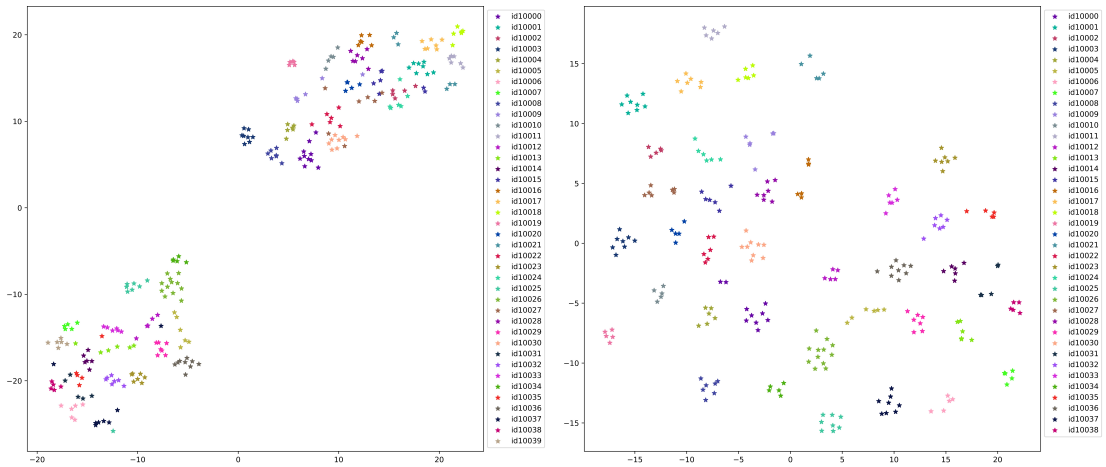
B. 纯语音特征的结果

我们从 VGGVox 子网络中提取特征，并训练了一个分类器以建立基线，识别性能达到 91%



(a) 语音片段的混淆矩阵　　(b) 从双分支网络提取的语音片段的混淆矩阵

图 3. 测试集数据的混淆矩阵：（左）从 VGGVox 网络提取的特征的混淆矩阵；（右）从所提出的、使用多模态数据训练的双分支网络提取的特征的混淆矩阵。混淆矩阵展示了 20 个身份的结果。（建议以彩色放大查看）



(a) 语音片段的可视化　　(b) 从双分支网络提取的语音片段的可视化

图 4. 测试集数据的 T-SNE 图：（左）从 VGGVox 网络提取的语音特征的 T-SNE 图；（右）从所提出的、使用多模态数据训练的双分支网络提取的语音特征的 T-SNE 图。选取了 30 个随机身份的特征。

。图 3a 显示了基线结果的混淆矩阵。此外，图 4a 中样本测试集的随机邻域嵌入显示，网络将多个类别的特征分布得相距很远，这降低了分类器基于这些特征的准确率。

C. 采用面部信息辅助的结果

表II展示了在从双分支网络中提取的特征上训练的分类器的结果。双分支融合特征的混淆矩阵见图3b。此外，图4b中样本测试集的T-SNE 图表明，该网络能够更有效地分布特征，使同一类别的特征彼此更接近，从而有助于支持向量机更好地学习说话人识别任务。

实验表明，在特征提取过程中，借助人脸信息辅助语音信号，能显著提升说话人识别效果。缺少面部信息时，系统容易受到噪声影响。当人脸信息辅助语音信息时，一种模态中退化的信息可以通过另一种模态恢复。

表II. VoxCeleb1 上的说话人身份识别性能。（越高越好）

方法	Loss	Top-1 %
I-向量 + 概率线性判别分析 + 支持向量机 [2]	-	60.8
卷积神经网络 [2]	-	80.5
VGGVox（基线）	-	91.0
网络 A（维度 =128） [28]	中心+Softmax	84.6
网络 B（维度 =128） [28]	中心+Softmax	89.5
Ours	交叉熵损失	**97.2**

五、结论

在本研究中，我们提出，多模态信息的存在能够提升说话人识别任务的性能。我们提出了双分支网络，用于从

面部与语音信号中提取特征。我们使用支持向量机，根据单一领域和多领域的特征对说话人进行分类。当我们将面部信息和语音信息同时作为模型的输入时，取得了令人鼓舞的结果。与仅利用单模态的 VGGVox 相比，我们的方法实现了更高的识别性能。此外，使用我们的模型提取特征时，同时输入语音和面部信号，效果优于仅输入说话人信息。这清楚地表明，面部信息有助于说话人识别。此外，面部信息带来的说话人识别性能提升也让我们认识到，一个人的面部与声音之间存在某种关联。另一项非常重要的贡献是，本研究为其他模态的分类和检索任务开辟了研究方向。

致谢。 作者衷心感谢群体机器人实验室和 NCRA 为我们的实验提供必要的设备与资源。

参考文献

[1] Jiquan Ngiam、Aditya Khosla、Mingyu Kim、Juhan Nam、Honglak Lee 和 Andrew Y Ng，《多模态深度学习》，载于《第*28*届国际机器学习大会（*ICML-11*）论文集》，2011年，第689–696页。[2] A. Nagrani、J. S. Chung 和 A. Zisserman，《VoxCeleb：大规模说话人识别数据集》，载于*INTERSPEECH*，2017年。[3] Joon Son Chung、Arsha Nagrani 和 Andrew Zisserman，《VoxCeleb2：深度说话人识别》，*arXiv*预印本 *arXiv:1806.05622*，2018年。[4] Ken Chatfield、Karen Simonyan、Andrea Vedaldi 和 Andrew Zisserman，《细节中的魔鬼归来：深入研究卷积网络》，*arXiv*预印本 *arXiv:1405.3531*，2014年。[5] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun，《用于图像识别的深度残差学习》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2016年，第770–778页。[6] Shah Nawaz、Muhammad Kamran Janjua、Ignazio Gallo、Arif Mahmood 和 Alessandro Calefati，《用于音频与视觉信号跨模态映射的深度潜在空间学习》，载于《*2019*年数字图像计算：技术与应用（*DICTA*）》。IEEE，2019年，第1–7页。[7] Arsha Nagrani、Samuel Albanie 和 Andrew Zisserman，《看见声音，听见面孔：跨模态生物特征匹配》，载于《*IEEE*计算机视觉与模式识别会议论文集》，2018年，第8427–8436页。[8] Yandong Wen、Mahmoud Al Ismail、Weiyang Liu、Bhiksha Raj 和 Rita Singh，《用于声音与人脸跨模态匹配的不相交映射网络》，载于《第7届国际学习表征会议（*ICLR 2019*），美国，*2019*年*5*月*6*日至*9*日》，2019年。[9] Shah Nawaz、Muhammad Saad Saeed、Pietro Morerio、Arif Mahmood、Ignazio Gallo、Muhammad Haroon Yousaf 和 Alessio Del Bue，《跨模态说话人验证与识别：多语言视角》，载于《*IEEE/CVF*计算机视觉与模式识别会议论文集》，2021年，第1682–1691页。[10] Arsha Nagrani、Samuel Albanie 和 Andrew Zisserman，《可学习的别针：用于个人身份的跨模态嵌入》，载于《欧洲计算机视觉会议（*ECCV*）论文集》，2018年，第71–88页。[11] Samuel Albanie、Arsha Nagrani、Andrea Vedaldi 和 Andrew Zisserman，《利用野外跨模态迁移进行语音情绪识别》，载于《第*26*届*ACM*国际多媒体会议论文集》，2018年，第292–301页。[12] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisserman，《对话：深度视听语音增强》，*arXiv*预印本 *arXiv:1804.04121*，2018年。[13] O. Wiles、A.S. Koepke 和 A. Zisserman，《从视频中自监督学习面部属性嵌入》，载于英国机器视觉会议，2018年。[14] Miyuki Kamachi、Harold Hill、Karen Lander 和 Eric Vatikiotis-Bateson，《让声音与面孔相配：跨模态身份匹配》，《当代生物学》，第13卷，第19期，第1709–1714页，2003年。[15] Sandra Pruzansky，《用于自动说话人识别的模式匹配程序》，《美国声学学会期刊》，第35卷，第3期，第354–358页，1963年。[16] Richard J Mammone、Xiaoyu Zhang 和 Ravi P Ramachandran，《鲁棒说话人识别：基于特征的方法》，《*IEEE* 信号处理杂志》，第13卷，第5期，第58页，1996年。[17] Umit Yapanel、Xianxian Zhang 和 John HL Hansen，《真实车载环境中的高性能数字识别》，载于第七届国际口语处理会议，2002年。[18] John HL Hansen、Ruhi Sarikaya、Umit Yapanel 和 Bryan Pellom，《噪声环境下的鲁棒语音识别：使用SPINE语料库进行评估》，载于第七届欧洲语音通信与技术会议，2001年。[19] Patrick Kenny，《说话人与会话变异性的联合因子分析：理论与算法》，*CRIM*，蒙特利尔，（报告）*CRIM-06/08-13*，第14卷，第28–29期，第2页，2005年。[20] Muhammad Saad Saeed、Muhammad Haris Khan、Shah Nawaz、Muhammad Haroon Yousaf 和 Alessio Del Bue，《通过融合与正交投影改进人脸-语音关联》，载于*ICASSP 2022——2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第7057–7061页。[21] Muhammad Saad Saeed、Shah Nawaz、Muhammad Haris Khan、Sajid Javed、Muhammad Haroon Yousaf 和 Alessio Del Bue，《学习用于人脸-语音关联的分支融合与正交投影》，*arXiv*预印本 *arXiv:2208.10238*，2022年。[22] Changil Kim、Hijung Valentina Shin、Tae-Hyun Oh、Alexandre Kaspar、Mohamed Elgharib 和 Wojciech Matusik，《关于学习人脸与声音的关联》，载于亚洲计算机视觉会议。施普林格，2018年，第276–292页。[23] Elad Hoffer 和 Nir Ailon，《使用三元组网络进行深度度量学习》，载于《基于相似性的模式识别：第三届国际研讨会，*SIMBAD 2015*，哥本哈根，丹麦，*2015*年*10*月*12*日至*14*日。论文集*3*》。施普林格，2015年，第84–92页。[24] Karen Simonyan 和 Andrew Zisserman，《用于大规模图像识别的超深卷积网络》，*arXiv*预印本 *arXiv:1409.1556*，2014年。[25] Yusuf Aytar、Carl Vondrick 和 Antonio Torralba，《SoundNet：从无标签视频中学习声音表征》，《神经信息处理系统进展》，第29卷，2016年。[26] Omkar M Parkhi、Andrea Vedaldi 和 Andrew Zisserman，《深度人脸识别》，2015年。[27] Zhilu Zhang 和 Mert Sabuncu，《用于训练含噪声标签深度神经网络的广义交叉熵损失》，《神经信息处理系统进展》，第31卷，2018年。[28] Sarthak Yadav 和 Atul Rai，《学习用于说话人识别与验证的判别性特征》，载于*INTERSPEECH*，2018年，第2237–2241页。