arXiv:2304.05754v1 [cs.SD] 12 Apr 2023

基于簇感知 DINO 的自监督学习：实现高性能鲁棒说话人验证

Bing Han，学生会员，*IEEE*，Zhengyang Chen，学生会员，*IEEE*，以及 Yanmin Qian，高级会员，*IEEE*

摘要—利用大规模人工标注数据集的深度学习方法，自动说话人验证任务取得了巨大成就。然而，收集大量标注良好的数据用于系统构建非常困难且成本高昂。近年来，由于不依赖有标签数据，自监督说话人验证受到了广泛关注。在本文中，我们提出了一种新颖且先进的自监督学习框架，无需使用任何有标签数据即可构建性能卓越的说话人验证系统。为避免基于对比学习的自监督学习中假负样本对的影响，我们采用无标签自蒸馏（DINO）框架作为初始模型，该模型无需利用负样本对即可进行训练。随后，我们进一步为 DINO 引入簇感知训练策略，以提升数据的多样性。在迭代学习阶段，由于无监督聚类会产生大量不可靠标签，伪标签的质量对系统性能至关重要。这促使我们提出动态损失门控和标签校正（DLG-LC）方法，以缓解不可靠标签导致的性能下降。具体而言，我们使用高斯混合模型（GMM）对损失分布进行建模，并动态获得损失门控阈值，以区分可靠标签和不可靠标签。此外，我们采用模型预测来校正不可靠标签，从而更好地利用不可靠数据，而不是直接丢弃它们。进一步地，我们将 DLG-LC 从单模态扩展到音视频数据集上的多模态，以进一步提升性能。实验在广泛使用的 Voxceleb 数据集上进行。与当前已知性能最佳的自监督说话人验证系统相比，即使使用更少的迭代次数、更小的模型和更简单的聚类方法，我们提出的方法在 Vox-O、Vox-E 和 Vox-H 测试集上分别取得了 22.17%、27.94% 和 25.56% 的相对等错误率提升。更重要的是，新提出的自监督学习系统在 Voxceleb 数据集上取得了与全监督系统相当的结果，但未使用任何人工标注数据。

索引词—**自监督说话人验证、聚类感知 DINO、动态损失门控、标签校正、多模态**

I. 引言

SPEAKER 话人验证（SV）是一项利用语音作为生物特征来验证说话人身份的任务。近年来，深度学习方法已广泛应用于说话人验证（SV）任务，并在多方面进行了探索，例如采用不同的模型架构 [2], [3], [4], [5], [6],训练目标 [7], [8], [9], 池化方法 [10], [11] 和等等，以取得优于高斯混合模型-通用背景模型（GMM-UBM） [12], i-vector [13]等传统方法的出色性能。然而，所有这些方法都基于全监督训练，通常需要大量带有准确人工标注的训练数据；但众所周知，收集大规模且标注完善的数据实际上十分困难，成本也很高。

为了降低对有标签数据的高度依赖，近年来自监督学习引起了广泛关注，一些研究者正致力于将其应用于说话人验证任务。受语音预训练模型在自动语音识别（ASR）任务中取得的巨大成功（例如 wav2vec 2.0 [14]和 HuBERT [15], ）的启发，一些研究者 [16]尝试提取通用语音表征，并直接在说话人验证任务上进行微调。由于这些预训练模型的训练未使用明确的说话人信息，直接微调的结果并不理想。在研究 [17],中，研究者探索利用从大规模无标签数据中学习到的语音表征来替代声学特征，然后照常训练常规的有监督深度模型。尽管取得了颇具前景的性能，但该方法仍需要有标签数据进行训练，而且由于预训练模型规模庞大，其参数量对于实际应用而言难以接受。

为充分利用大规模无标签数据，受文本转语音（TTS）任务的启发， [18]中研究了一种生成式方法，借助音素信息分离说话人表征。随后，一些研究人员提出了一种假设：从同一话语中截取的语音片段属于同一说话人，而从不同话语中截取的片段则属于不同说话人；这一假设大致成立。基于该假设，研究人员 [19], [20], [21], [22], [23] 通过对比学习最大化同一话语中不同片段之间的信息，致力于获得判别式说话人表征。之后，受 [24], 启发，研究人员提出了 [25]迭代学习框架，以进一步提升自监督说话人验证系统的性能。这种先进系统通常包含两个阶段。阶段 I 中，采用基于对比学习的目标函数训练说话人编码器。阶段 II 中，系统使用阶段 I 的预训练模型，通过聚类估计伪标签，并将其作为监督信号来训练新的编码器。该过程通过迭代执行，持续提升系统性能。

这一两阶段框架已取得出色的性能 [26], [27], [28], [29], [30],，但仍存在许多不足之处，

部分研究结果已发表于 Interspeech 2022 [1]。所有作者均来自上海交通大学人工智能研究院、教育部人工智能重点实验室、计算机科学与工程系的 X-Lance 实验室，地址：中国上海市，邮编 200240（电子邮箱：{hanbing97, zhengyang.chen, yanminqian}@sjtu.edu.cn）

这些不足限制了系统性能的进一步提升。对于阶段 I 中的对比学习方法，从不同话语中裁剪出的语音片段被视为负样本对，并在说话人空间中彼此推远。然而，在实际情况下，不同话语可能属于同一说话人，这表明这一不准确的假设可能会导致错误。对于第二个迭代阶段， [24], [25]研究已证明，聚类算法生成的许多伪标签缺乏可靠性，可能会使模型产生混淆并导致性能下降。因此，提高性能的关键在于找到一种筛选高质量伪标签的方法。基于这一假设，在 [29],他们观察到，损失较低的数据比带有不可靠标签的数据更可靠，随后提出了一种损失门控学习策略，通过设置损失阈值来区分可靠标签与不可靠标签。只有损失低于阈值的数据才能用于更新网络。尽管这种方法进一步提升了性能，但每次迭代中手动设置的阈值缺乏灵活性，而且带有不可靠标签的数据未得到充分利用。

在本文中，我们提出了几种用于自监督说话人验证的新策略。首先，在第一阶段预训练中，我们引入了 DINO（无标签蒸馏） [31] ，该方法仅基于最大化从同一话语中采样的增强片段对之间的相似度。为了最小化信道和环境影响并增加数据的多样性，我们提出了一种面向聚类的 DINO 训练策略，以进一步提升其性能。在第二个迭代阶段，我们使用包含两个组分的高斯混合模型对损失分布数据进行建模，其中每个组分分别表示具有可靠标签或不可靠标签的数据。然后，使用通过估计的高斯混合模型计算得到的动态损失门控（DLG）阈值来区分这两类标签，这比手动调节的阈值更加灵活。此外，受半监督学习工作 [32], [33], [34], 的启发，我们提出了标签校正（LC），利用模型的预测作为目标标签，并使用它来校正不可靠的伪标签，而不是直接丢弃不可靠数据 [29]。最后，我们将多模态引入上述提出的 DLG-LC策略和聚类步骤中。得益于不同模态之间互补的音频和视觉信息，DLG-LC 能够更有效地选择具有可靠标签的数据。

本文的主要贡献总结如下：

1) 引入 DINO 框架作为自监督学习框架，以获得初始预训练模型。该框架无需负样本对，可避免假负样本对的影响。此外，我们设计了簇感知训练策略来增强 DINO，从而提升数据多样性并取得更好的性能。2) 为了在第二个迭代阶段更有效、灵活地筛选高质量数据，我们提出了动态损失门控（DLG），可动态确定损失门控阈值，以筛选带有可靠标签的数据。同时，我们还采用了标签校正（LC），以进一步改善结果。

3）DLG-LC 方法进一步从音频单模态扩展到视听多模态。多模态数据利用多模态知识，使可靠标签选择更加高效。

4) 采用这些策略后，与当前自监督学习领域最先进水平（SOTA）的系统相比，我们取得了显著的性能提升，而且迭代次数更少、模型更小、聚类方法更简单。更值得期待的是，这一新提出的自监督学习框架可达到当前全监督系统的 SOTA 水平，并取得相当的性能。

II. 用于说话人验证的自监督学习

本节回顾了常用的两阶段自监督说话人验证框架，包括针对预训练模型的第一阶段对比学习阶段，以及第二阶段迭代学习阶段。

A. 基于对比学习的自监督说话人验证

自监督学习 (SSL) 是一种无监督训练方式，可以设计前置任务或代理任务，并从数据本身学习表示。常见的 SSL 方法大致可分为两类：生成式 [18] 和对比式 [35], [36]方法。基于这样的假设：从同一话语中采样的片段属于同一说话人，而来自不同话语的片段则属于不同说话人，大多数说话人验证任务研究都聚焦于对比学习方法。其中，SimCLR [37]是最流行的对比学习框架之一。其基本思想是缩小从同一话语裁剪出的增强片段的表示之间的距离，同时增大来自不同话语的负样本对之间的距离。此外，MoCo [38] 框架通过包含队列的动态字典和移动平均编码器进一步提升了性能。基于这些框架，人们提出了许多方法，例如均衡学习 [23],、增强对抗训练 [20],、信道不变训练 [21], 和原型动量 [22] ，以学习更具判别性的说话人表征。

B. 自监督说话人验证的迭代框架

考虑到对比学习的假设可能会自然引入标签错误并导致模型性能下降， [25], 中提出了一种迭代式自进化框架，以进一步提升自监督说话人验证系统的性能。该框架主要分为两个阶段，具体如下：

- 阶段 I：预训练 1) 使用对比学习或其他自监督学习方法预训练说话人编码器，作为初始模型。 2) 使用预训练模型提取训练集的说话人嵌入，然后应用聚类算法分配伪标签。

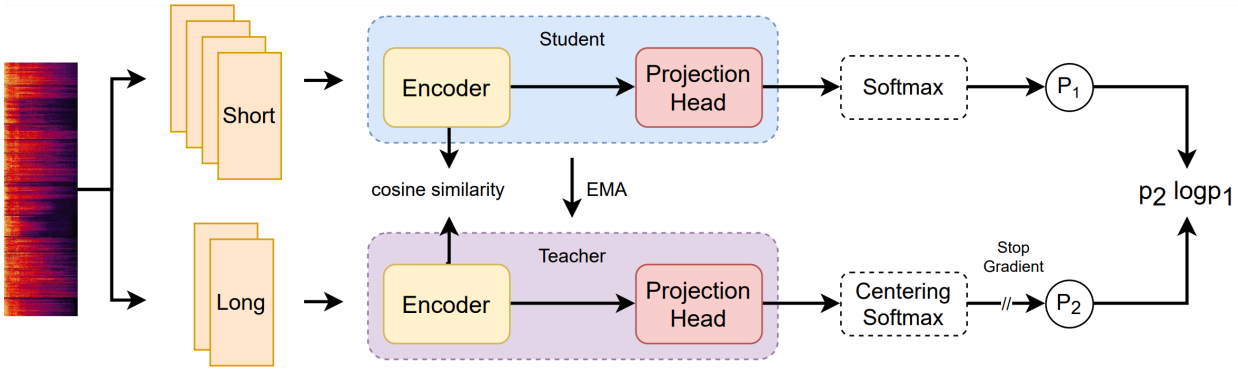

图 1：用于自监督说话人表征学习的无标签蒸馏（DINO）框架

- 阶段 II：迭代训练与伪标签。
  1) 使用上一步生成的伪标签训练新的编码器。2) 使用新的编码器执行聚类算法，以更新伪标签。3) 重复阶段 II 多次，直到模型收敛。

尽管由于多次迭代，该框架需要较高的计算资源，但它被广泛用于 [26], [27],[39], [28], [29]，因为其先进的性能。此外，该框架在 [30]中扩展到视听数据集，并借助聚类算法中的多模态信息实现了更好的性能。

III. 用于说话人验证的簇感知 DINO

对于先前工作中的基于对比学习的方法，它们共享一个相同的假设，即一个批次中的片段属于不同的说话人。但这一假设并不总是成立，因为重复的说话人可能会出现在同一个批次中。以 VoxCeleb2 上的统计数据为例，我们可以通过公式计算 VoxCeleb2 上重复说话人的概率。1结果列于表 I 中。I。

$$p_{repeat}(S,N) = 1 - \frac{A_S^N}{S^N} = 1 - \frac{S!}{S^N(S-N)!} \quad (1)$$

其中，S 是训练集中的说话人数量，N 是批量大小。

表 I：一个批次中重复说话人的概率

批量大小	16	32	64	128	256
概率	0.020	0.080	0.286	0.745	0.996

根据表格，批量越大，重复出现的概率越高，这会对模型产生不良影响。我们可以使用较小的批量来缓解这个问题，但这会降低性能 [38]。

A. 基于 *DINO* 的自监督学习

为解决这一问题，我们将无负样本对 DINO [31] 引入自监督说话人验证任务，整体框架如图 1所示。

首先，使用 a 从一个话语中随机采样 4 个短 $\{x_1^s, x_2^s, x_3^s, x_4^s\}$ 和 2 个长片段$\{x_1^l, x_2^l\}$ 多裁剪策略 [40], 并且可以使用较长的片段来提取更稳定的说话人嵌入。值得注意的是，在采样时，这些片段应尽可能少地重叠。与之前的工作相同 [20], [21], [22], [23], 我们仍然遵循这样的假设：从同一话语中裁剪出的片段属于同一个说话人，然后通过添加噪声或房间脉冲响应，对它们应用不同种类的数据增强，以实现鲁棒性能。不同于 SimCLR [37], 仅使用一个编码器进行对比学习，我们的模型不仅包含一个学生 编码器，还包含一个动量 教师 编码器，其架构类似于知识蒸馏 [41]。增强后，所有片段都通过学生编码器，而只有较长的片段通过教师编码器，从而通过最小化短片段到长片段之间的对应关系的交叉熵 $H(\cdot)$来鼓励两个分布之间的对应关系，如以下公式所示。2:

$$L_{ce} = \sum_{x \in \{x_1^l, x_2^l\}} \sum_{x' \in \{x_1^l, x_2^l, x_1^s, \ldots, x_4^s\}} H(P_t(x) \mid P_s(x')) \quad (2)$$

其中，动量教师网络 f_{θ_t}和学生网络 f_{θ_s} 的输出分布分别记为 P_t 和 P_s。 P 可通过使用 softmax 函数对输出进行归一化计算得到:

$$P_s(x) = Softmax(\frac{f_{\theta_s}(x)}{\epsilon_s}) \quad (3)$$

其中，$\epsilon_s > 0$ 是温度参数，可用于控制输出分布的锐化程度。同样，P_t 也有相应的公式，其温度为 $\epsilon_t > 0$。此外，使用批次均值对教师模型的输出分布进行中心化。训练过程中，同时进行锐化和中心化，以避免平凡解 [31]。

教师和学生具有相同的架构，但由于更新方法不同，参数也不同。学生通过梯度下降进行更新，而教师则通过对学生的参数进行指数移动平均（EMA）来更新。EMA 的更新规则为:

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s \quad (4)$$

其中，训练期间 λ 按照余弦调度 [42] 从 0.996 调整至 1。编码器提取说话人嵌入。随后，将说话人嵌入输入投影头，该投影头包含一个具有隐藏层的三层感知机

维度为 2048，后接 ℓ_2 归一化层和一个具有K 维度的权重归一化全连接层。整个架构与 [40].

此外，还加入了基于余弦的 consistency 损失，以确保说话人嵌入被编码到更适合后续评分和聚类的余弦空间中。该方法通过最大化从同一说话人提取的嵌入之间的余弦相似度来发挥作用。最后，总损失可概括为系数 α:

$$L_{dino} = L_{ce} + \alpha \sum_{e \in \{e_1^l, e_2^l\}} \sum_{e' \in \{e_1^l, e_2^l, e_1^s, \ldots, e_4^s\}} \left(1 - \frac{e \cdot e'}{\|e\| \, \|e'\|}\right) \quad (5)$$

其中 e 表示从编码器中提取的说话人嵌入。

B. DINO 上的聚类感知训练

对于传统DINO，所有片段都从同一段话语中采样，以形成正样本对。受话语时长限制，这些片段通常具有很大程度的重叠。如上所述，DINO 的优化通过最小化正样本对的两个分布之间的交叉熵，来鼓励短到长对应关系。由于片段中存在大量重叠部分，模型可能倾向于更加关注重叠部分的内容、信道及其他无关信息，而忽略音频中的说话人信息。尽管我们会对片段添加不同类型的数据增强，但数据仍然缺乏多样性，这可能会导致模型优化方向错误。

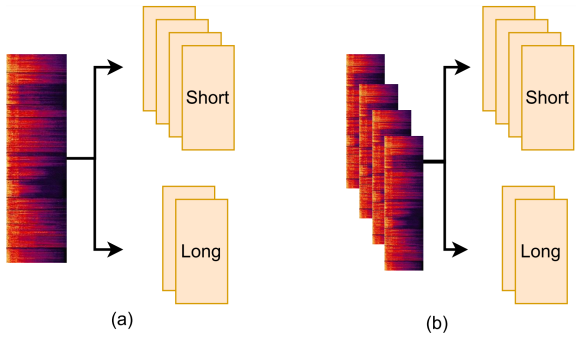

图 2：传统 DINO 与聚类感知训练 DINO 的区别。(a) 传统 DINO：从同一话语中采样长、短片段，组成正样本对。(b) 聚类感知训练 DINO：通过简单的聚类算法，我们认为同一聚类中的同一说话人具有相同身份，并从相应的聚类中裁剪片段。

为了减少片段之间的重叠并增加数据的多样性，我们提出了一种面向聚类的（CA）DINO 训练策略，同时尽可能保持原有假设，以下将其命名为 CA-DINO。我们将模型训练分为两个阶段。在训练的早期阶段，我们根据传统DINO 策略优化模型。当模型具备提取判别性说话人表征的能力后，训练过程将进入下一阶段。在第二阶段，使用提取的说话人嵌入执行聚类算法，并且我们假设同一聚类中的话语属于同一个人。如图所示2，正样本对是从属于同一聚类的多个话语中采样得到的，而不是来自单个话语。这些样本对来自同一个说话人，但具有不同的语音内容和信道，因此具有较高的数据多样性，并使模型更加关注说话人的信息，而非无关因素。考虑到提取嵌入说话人的资源消耗，聚类操作将每隔几轮执行一次。

IV. 使用动态损失门控和标签校正进行迭代学习

基于所提出的 CA-DINO 自监督学习，我们进一步应用迭代学习框架 [25] 来提升自监督说话人验证的性能。在迭代过程中，一个严重的问题是生成的伪标签含有大量噪声，会使网络产生混淆并导致性能下降。考虑到这一局限，已有多项工作致力于筛选高质量的伪标签。在 [25],中，研究者采用了一种激进的训练方法，利用聚类置信度来净化标签，但收益有限。在 [29], 中，研究者进行了一个简单实验，发现损失较低的数据更可靠。随后，他们提出损失门控（LG）策略，通过设置固定阈值来筛选损失较低的数据，并仅使用这些数据更新模型。采用 LG 策略后，系统性能明显提升，但该方法中的阈值设置严重依赖人工经验，且不可靠数据未得到充分利用。

在本节中，我们将介绍所提出的 DLG-LC，以动态调整损失门控阈值并纠正不可靠的伪标签，从而充分利用数据；随后，我们将此 DLG-LC 方法扩展至多模态，以进一步提升性能。

A. 动态损失门控

为了确定合适的损失门控阈值，我们实现了LG学习，并对Voxceleb上的损失值分布进行了可视化分析 2 [43] 数据集。损失值的直方图见图3。从图中可以明显看出，分布中存在两个锐化峰值。在 [44] 中进行的类似实验表明，具有可靠标签和不可靠标签的数据分别可以由两个峰值表示。如果我们能找到一种对分布进行建模的方法，就可以随着损失分布的变化动态确定损失门控阈值，从而避免费力的手动调参。

高斯分布是一种重要的连续型实值随机变量概率分布，其概率密度函数的一般形式定义于公式中。6。

$$\mathcal{N}(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \quad (6)$$

其中，位置参数和尺度参数分别用μ 和 σ 表示。高斯分布的形状类似于

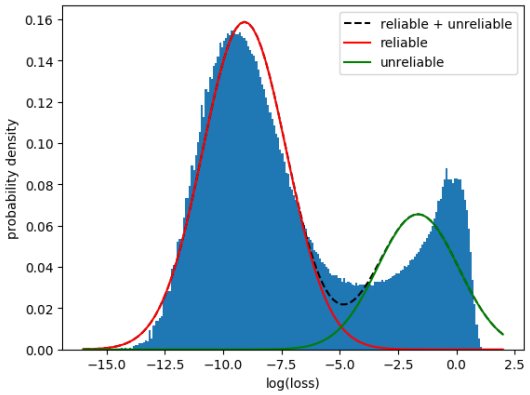

图 3：损失门控（LG）学习的损失分布 [29]在 VoxCeleb 上 2 [43]。损失值按对数函数缩放，曲线由包含两个成分的高斯混合模型估计。

钟形曲线，两侧低、中间高，与图3中的损失“峰值”非常相似。在这种情况下，可以采用具有两个分量的高斯混合模型（GMM），分别对可靠样本和不可靠样本的损失分布进行建模：

$$p(x) = \lambda_1 \mathcal{N}\left(\mu_1, \sigma_1^2\right) + \lambda_2 \mathcal{N}\left(\mu_2, \sigma_2^2\right) \tag{7}$$

其中 λ_1 和 λ_2 表示两个高斯分量的权重。拟合后，拟合曲线如图 3所示。显然，两个加权高斯分量可以用来拟合这两个“峰值”。随后，通过计算两个分量对应概率相等时的损失值，即可轻松获得损失门控阈值 τ_1，从而区分可靠数据和不可靠数据：

$$\tau_1 : p_1(\tau_1) = p_2(\tau_1) \tag{8}$$

其中 $p_1(x) = \lambda_1 \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ 和 $p_2(x) = \lambda_2 \mathcal{N}\left(\mu_2, \sigma_2^2\right)$。对于每个训练轮次，所有损失值都会被记录下来，以重新估计高斯混合模型的参数，因此 τ_1 可以根据当前训练状况进行动态调整。

我们的 DLG 引入了这一动态损失门控阈值τ_1 到说话人分类损失函数 ArcMargin Softmax (AAM) 中 [45] 以筛选数据，只有损失低于该阈值的保留数据才会用于更新网络的参数。

$$L_{DLG} = \sum_{i=1}^{N} \mathbb{1}_{l_i < \tau_1} \log \frac{e^{s(\cos(\theta_{y_i,i}+m))}}{Z} \tag{9}$$

其中 $Z = e^{s(\cos(\theta_{y_i,i}+m))} + \sum_{j=1,j\neq i}^{c} e^{s(\cos(\theta_{y_i,i}))}$, $\theta_{j,i}$ 是列向量 W_j 与嵌入 x_i之间的夹角。s 是缩放因子， m 是用于控制间隔的超参数。AAM 可以增大最近说话人之间的间隔，因此被广泛用于说话人识别任务。

B. 标签校正

对于那些损失较大的不可靠数据，直接将其丢弃是浪费的。因此，我们提出标签校正（LC）策略，动态校正伪标签，从而

算法 1： 所提出的动态损失门控以及标签校正

输入： 小批次 $D_m = \{(x_1, \ x_2, \ y)\}_{i=1}^{n}$；两个阈值 τ_1 和 τ_2; 网络 $g(\cdot)$; 锐度因子 ϵ_c

输出： 小批次的损失

对于 $(x_1, x_2, y) \in D_m$ **执行**
  $x_{clean}, x_{aug} = x_1$, 增强(x_2) # 增强一个片段
  $p_{clean}, p_{aug} = g(x_{clean}), \ g(x_{aug})$ # 输出分布
  根据 AAM-softmax 损失 l_{clean} 和 l_{aug} 计算。伪标签 y
  记录 l_{clean} 值
  如果 $l_{clean} < \tau_1$ **那么**
    返回 l_{aug} # 伪标签 y 是可靠的
  else
    如果 最大值$(p_{clean}) > \tau_2$ **则**
      $\hat{p}_{clean}$ =锐化(p_{clean}, ϵ_c) # 锐化分布
      计算交叉熵 l $\hat{p}_{clean}$ 与 p_{aug}之间
      返回 l
    else
      返回 0 # 预测不可靠
    结束 if
  结束 if
结束 for
经过一个训练轮次后，根据记录的损失值重新估计高斯混合模型然后更新 τ_1

从而有效利用不可靠数据。研究者在 [46]中指出，网络能够将噪声样本聚类到其正确类别。为了发挥这一能力，我们假设模型输出的预测比通过聚类生成的伪标签更加可靠。因此，将预测后验概率视为目标标签，并纳入目标损失函数，以防止模型拟合不准确的标签。然而，并非所有预测标签都适合用于训练。受 [47], [33], 启发，我们假设，如果模型为某个可能类别分配了较高的概率，则该预测标签具有较高的置信度。随后，引入另一个阈值 τ_2 ，保留最大类别概率高于τ_2的预测，并将标签校正损失定义为如下公式 10:

$$L_{LC} = \sum_{i=1}^{N} \mathbb{1}_{l_i > \tau_1, \max(\hat{p_i}) > \tau_2} H(\hat{p_{clean}} \mid p_{aug}) \tag{10}$$

其中， p_{aug} 表示增强片段的输出概率，$\hat{p}_{clean}$ 表示其对应的干净版本。 $H(\cdot)$ 此处表示两个概率分布之间的交叉熵损失函数。此外，为了鼓励分布更加尖锐，对$\hat{p}_{clean}$ 应用锐化操作，锐度因子为 ϵ_c ，如公式3所述。

然后，将 DLG 损失和 LC 损失结合起来，根据公式11优化说话人模型。

$$L = L_{DLG} + L_{LC} \tag{11}$$

更具体地说，文中详细给出了用于描述 DLG-LC 算法流程的伪代码，如算法1所示。

C. 融入多模态

研究人员在 [30] 中将多模态信息引入数据聚类步骤，以生成更

准确的自监督说话人验证伪标签。在我们的工作中，考虑到同一视频中的音频和视觉信息具有相同说话人身份这一特点，我们也尝试将视觉模态加入 DLG-LC 方法，以更好地利用数据，并希望取得进一步提升。我们主要从两个方面融合视觉信息：一是利用视觉信息帮助 DLG-LC 选择更可靠的数据；二是在数据聚类过程中改善聚类结果。

1) 基于多模态的 *DLG-LC*：与单模态 DLG-LC 不同，我们对可靠数据的筛选策略进行了少许调整。对于多模态数据，我们将使用两个独立的编码器分别对音频和视觉数据进行编码。然后，通过记录损失值，我们可以分别得到音频和视觉的两个损失门限。对于一个音频-视觉实例，只有当其损失值均低于这两个损失门限时，才能将其视为可靠标签；随后，我们将使用公式中定义的 AAM Softmax 对该实例进行优化。9.

对于不可靠数据，将对其进行多模态标签校正。首先，我们比较两个模态网络的预测标签是否一致。如果两个模型的预测属于同一类别，则说明预测准确率相对较高。与使用软标签进行训练的单模态标签校正不同，我们的输出经过多模态验证，因此可靠性更高。因此，我们使用“硬”标签（即模型分布的最大值索引）作为标签，通过 AAM Softmax 优化模型。如果网络对预测标签的判断不一致，则分别使用软标签优化模型。

2）基于多模态的数据聚类：在前一个训练步骤中，多模态信息仅用于筛选可靠数据，而两种模态的模型在结构上并无关联。因此，我们可以独立获得音频 g_a（·）和视觉编码器 g_v（·）。给定一个包含音频 x_a 和视觉模态 x_v的数据集，我们可以使用训练好的编码器分别提取音频嵌入 e_a 和视觉嵌入 e_v 。考虑到音频和视觉嵌入包含来自不同模态的互补信息，我们在联合表示 $e_{av} =$（e_a，e_v）上进一步进行聚类，该联合表示由音频嵌入和视觉嵌入拼接而成。通过联合操作，该表示将具有更强的判别性，聚类结果也将更加鲁棒。随后，在这些音频-视觉联合嵌入上通过 k-means 生成下一次迭代所需的伪标签。

V. 实验设置

A. 数据集

实验在 VoxCeleb 上进行 [48]，[43]，该数据集是用于说话人识别任务的大规模视听数据集。在自监督学习的阶段 I 和 II 中，我们采用 VoxCeleb 开发集 2 [43] 训练网络，此过程中不使用任何说话人身份信息。由于我们在迭代学习阶段引入了视觉特征，我们排除了数据集中缺少视频的一些话语。最终，视听训练集包含从 YouTube 提取的 5，994 名说话人的 1，091，251 条话语。

在评估中，我们报告了 [43] 中定义的 3 个测试集上的实验结果：VoxCeleb 原始、扩展和困难测试集。**Vox-O** 是 VoxCeleb 1 的原始测试集，包含来自 40 位说话人的 37，720 个试验。**Vox-E** 是一个试验列表，使用整个数据集构建，包含来自 1251 位说话人的 581，480 个试验。**Vox-H** 是一个困难评估列表，由从 VoxCeleb 1 中 1190 位说话人采样得到的 552，536 对组成，其中所有说话人均来自相同的国籍和性别。

B. 评估指标

本文采用的主要指标为：(i) 等错误率（EER），即接受率和拒绝率相等时的错误率；(ii) 归一化最小检测代价函数（最小检测代价函数），其定义见公式。12 ：

$$C_{det} = C_{miss} \times P_{miss} \times P_{tar} + C_{fa} \times P_{fa} \times (1 - P_{tar}) \quad (12)$$

其中，我们将先验目标概率 P_{tar} 设置为 0。01，并为漏检 C_{miss} 和误报 C_{fa}设置相等的权重。等错误率和最小检测代价函数都是说话人验证系统常用的评估指标。

C. 数据增强

1) 音频：为生成额外的训练样本并提高数据的多样性，我们通过以下方式执行在线数据增强策略： [49] 从 MUSAN 中添加背景噪声或卷积混响噪声 [50] 以及 RIR数据集 [51] ，分别进行增强。MUSAN中的噪声类型包括用于背景加性噪声的环境噪声、音乐、电视声和多人交谈噪声。我们可以通过将噪声与原始语音直接在时域波形中混合来获得增强数据，并随机应用 5至 20 dB之间的信噪比（SNR）。对于混响，则使用 40,000 个模拟房间脉冲响应（RIR）执行卷积运算 [51]。应用数据增强后，我们对波形值进行归一化，以实现稳定训练。我们使用 80 维对数梅尔滤波器组能量作为声学特征，其采用长度为 25ms 的汉明窗和 10 ms 的窗移；同时，我们的实验未涉及语音活动检测（VAD）。

2）视觉：对于 VoxCeleb 1 和 2 数据集中的每个视频片段，每秒提取一帧图像。随后，我们使用 MTCNN [52] 预测的关键点对提取帧中的人脸进行对齐，并使用相似变换将人脸区域映射为相同形状（$3 \times 112 \times 96$）。为了更好地提取视觉特征，我们将图像调整为模型最常用的尺寸（$3 \times 224 \times 224$）。接下来，以一定概率对原始图像应用多种数据增强策略，包括随机颜色失真、随机水平翻转、随机灰度变换和随机高斯模糊。最后，在将图像输入模型之前，我们将每张图像的像素值归一化到 [-0.5、0.5] 的范围内。

D. CA-DINO 设置

1）DINO：考虑到训练时间和内存限制，对于 DINO，我们采用 ECAPA-TDNN [53] 作为音频编码器，以学习具有判别性的说话人表征。ECAPA-TDNN 是一种时延神经网络（TDNN） [3]骨干网络，突出强调通道注意力、传播和聚合。它采用通道与上下文相关的注意力机制——多层特征聚合（MFA） [54], ，以及挤压-激励（SE） [55] 和残差块。ECAPA-TDNN 的模型架构如表 III 所示。对于每段话语，随机裁剪两个较长（3 秒）和四个较短（2 秒）的片段，并将其视为正样本对。值得注意的是，所有片段都会进行数据增强，之后由编码器将其编码为 192 维说话人嵌入。与配置 [31]，类似，DINO 投影头中的 K 设为 65,536。教师ϵ_t 和学生 ϵ_s 的温度分别为 0.04 和 0.1。此外，我们将余弦损失权重 α 设为 1.0，以平衡两种损失。整个训练过程持续 150 个训练轮次。模型参数使用随机梯度下降（SGD）算法更新，权重衰减为 $5e$-5。学习率在前 20 个训练轮次中从 0 线性升至 0.2，随后按照余弦调度 [42]衰减至$1e$-5。此外，动量也按照余弦调度从 0.996 变化至 1.0。

2) 聚类感知训练：对于聚类感知训练策略，我们在前 90 个训练轮次中正常训练模型。之后，每 5 个训练轮次在整个训练集上应用基于 k-均值的聚类算法，该算法由 faiss 库 [56] 支持。聚类结果用于生成训练数据。正样本对从属于同一聚类的话语中采样，而不是仅从单个话语中采样。

表 II：视觉编码器 ResNet34 [57]的模型架构。C （核大小、通道）表示二维卷积层。 $[\cdot]$ 表示残差块，L是输入的图像尺寸。

层	结构	输出尺寸
输入	-	$3 \times L \times L$
二维卷积	C(3 × 3, 32)	$32 \times L \times L$
残差块 1	$\begin{bmatrix} \text{C}(3\times3,32)\ \text{C}(3\times3,32) \end{bmatrix} \times 3$, 步长 2	$32 \times \frac{L}{2} \times \frac{L}{2}$
残差块2	$\begin{bmatrix} \text{C}(3\times3,64)\ \text{C}(3\times3,64) \end{bmatrix} \times 4$, stride 2	$64 \times \frac{L}{4} \times \frac{L}{4}$
残差块 3	$\begin{bmatrix} \text{C}(3\times3,128)\ \text{C}(3\times3,128) \end{bmatrix} \times 6$, stride 2	$128 \times \frac{L}{8} \times \frac{L}{8}$
残差块 4	$\begin{bmatrix} \text{C}(3\times3,256)\ \text{C}(3\times3,256) \end{bmatrix} \times 3$, 步幅为 2	$256 \times \frac{L}{16} \times \frac{L}{16}$
嵌入	-	192

E. DLG-LC 设置

1) 单模态：在此阶段，为了与[29], 进行公平比较，我们也采用 ECAPA-TDNN [53] 作为音频编码器来提取说话人嵌入。对于聚类，我们选择k-均值算法为训练集分配伪标签。不同于一些直接将

表 III：音频编码器 ECAPA-时延神经网络 [53] 的模型架构。C(卷积核大小，信道)表示一维卷积层。 F 是输入声学特征的维度，由梅尔频谱图的频率箱数量决定。 T 与语音片段的帧数相关。

层	结构	输出尺寸
输入	-	$F \times T$
一维卷积	C(5, 512)	$512 \times T$
SE-Res2Block 1	C(1, 512) C(3,64) × 8，膨胀率 2 C(1, 512)	$512 \times T$
SE-Res2Block 2	C(1, 512) C(3,64) × 8，膨胀率 3 C(1, 512)	$512 \times T$
SE-Res2块 3	C(1, 512) C(3，64) × 8，膨胀率 4 C(1, 512)	$512 \times T$
一维卷积	C(1, 1536)	$1536 \times T$
池化层	注意力统计池化	3072×1
嵌入	-	192

将真实说话人的数量作为簇的数量，我们选择 7500 个簇来验证我们方法的鲁棒性。对于标签校正，锐化参数 ϵ_c 和阈值 τ_2 分别设为 0.1 和 0.5。学习率以指数方式从 0.1 衰减至 $5e$-5，动量和权重衰减分别设为 0.9 和 $1e$-4。最后，训练过程将持续 100 个训练轮次。

2）多模态： 对于基于音视频的 DLG-LC，除添加图像编码器外，其他配置均与单模态设置一致。我们采用 ResNet34 [57] 作为视觉编码器的骨干网络，这与近期研究 [58], [59]类似。更多细节见表II。

VI. 实验结果

实验分为六个部分。在第 VI-A节中，我们报告所提出的簇感知 DINO 在阶段 I 与先前工作的性能比较，并讨论簇的数量如何影响簇感知训练策略。在第 VI-B节中，我们报告在小规模有标签数据上微调 CA-DINO 后的说话人验证性能。在第 VI-C节中，我们通过消融研究展示所提出的 DLG-LC 的有效性。接着，第 VI-D 节和第 VI-E 节表明，所提出的动态损失门控和标签校正能够提升单模态和多模态场景下的性能。最后，在第 VI-F节中，我们对新提出的自监督学习方法与先前工作进行了全面比较，结果证明了我们系统的优越性和鲁棒性。

A. 基于 CA-DINO 的说话人验证评估

表 IV 报告了我们提出的方法与其他先前自监督说话人模型的说话人验证性能。所有方法均在 VoxCeleb 2 上训练，未使用任何说话人标签，并在 Vox-O 测试集上评估。根据结果可以发现，基于对比

表 IV：所提出的 CA-DINO 与其他自监督说话人验证方法的性能比较。SSL 表示自监督学习。在 Vox-O 测试集上评估等错误率（EER，%）和最小检测代价函数（minDCF，p=0.01）。

SSL 方法	等错误率 (%)	最小检测代价函数
Disent [35]	22.090	-
CDDL [36]	17.520	-
GCL [19]	15.260	-
i-vector [20]	15.280	0.63 (p=0.05)
AP + AAT [20]	8.650	0.45（p=0.05）
SimCLR +均匀 [21]	8.280	0.610
MoCo + WavAug [22]	8.230	0.590
Unif+CEL [23]	8.010	-
DINO	31.233	0.990
+ 指数移动平均	4.404	0.434
+ + 簇感知（CA）	**3.585**	**0.353**

学习 [20], [21], [22], [23] 的方法与传统方法相比，性能有了大幅提升 [35], [36], [19]。我们提出的无需负样本对的 CA-DINO 再次实现了显著的性能飞跃，这表明负样本对确实是性能提升的瓶颈。此外，我们还在表IV底部给出了 CA-DINO 的消融研究。训练不使用指数移动平均（EMA）的 DINO 时，模型难以收敛，结果也很差，这表明 EMA 是防止模型坍塌的关键。随后，我们在训练 DINO 时采用聚类感知（CA）策略，性能得到进一步提升。所提出的 CA-DINO 达到了 **3.585%**的等错误率，较此前已发表的自监督说话人验证最佳性能 [23]相对提升了 55.24%。

表V：采用不同聚类数量的聚类感知训练性能比较。等错误率（%）在Vox-O 测试集上进行评估。此处的 1080k 表示一个话语为一个类别，这等同于不采用聚类感知策略进行训练。

# 簇	1080k	30k	20k	10k	5k
等错误率(%)	4.404	3.909	3.946	**3.585**	3.978

在簇感知训练过程中，存在一个 k-均值聚类操作。我们还进行了实验，以探究簇数量对性能的影响，结果见表V。结果表明，与基线系统（1080k）相比，我们提出的簇感知训练策略在所有给定的簇数量下都能带来显著且稳定的提升。同时，簇数量为 10k 的 CA-DINO 优于其他系统，这说明合理设置簇数量可以最大限度地提升性能。

B. 使用有标签数据的预训练-微调框架对 *CA-DINO* 的评估

为更好地展示我们提出的 CA-DINO 的卓越性能，我们探索了自

表VI：使用 VoxCeleb 1 中不同数量的有标签数据对预训练自监督模型进行微调的等错误率(%)对比。评估结果基于 Vox-O，该数据集是 VoxCeleb 1 的测试集。

初始模型	None	10%	20%	50%	100%
随机	32.78	6.893	5.276	3.691	2.755
SimCLR	8.547	4.388	3.797	3.266	2.936
CA-DINO	**3.585**	**2.393**	**2.356**	**2.016**	**1.835**

监督学习与预训练-微调框架，即在下游说话人验证任务中，使用少量有标签数据对自监督模型进行微调。我们从 VoxCeleb 中随机抽取 10%/20%/50%/100% 的有标签话语1 [48] 作为监督数据，并使用这些数据对自监督模型进行微调。

从表 VI可以看出，与从头开始训练的模型相比，自监督模型（包括 SimCLR 和本文提出的 CA-DINO）均取得了显著提升，这表明在低资源条件下，采用初始化效果更好的预训练模型至关重要。此外，将本文提出的 CA-DINO 与 SimCLR 进行比较可知，本文提出的非对比式 CA-DINO 明显优于 SimCLR，并且仅使用少量有标签数据，就能在下游说话人验证任务中取得良好表现。此外，仅使用 10% 的有标签数据时，CA-DINO 的表现甚至优于全监督系统，即 2.393% 对比 2.755%，这有助于节省大量人工标注工作。

C. 对所提出的 *DLG-LC* 的评估

表VII：在第1次迭代中，所提出的 DLG-LC 在 Vox-O、Vox-E 和 Vox-H 上的等错误率（%）对比。本实验中的伪标签由我们预训练的 CA-DINO 系统估计。SimCLR 和 CA-DINO 在此表示训练系统时使用所有带有估计伪标签的数据作为监督信号，未采用任何数据选择策略。

方法	阈值	Vox-O	Vox-E	Vox-H
SimCLR	-	6.281	7.428	11.54
CA-DINO	-	**2.909**	**3.315**	**5.692**
CA-DINO				
+ LG [29]	1	**2.441**	**2.930**	**4.892**
+ LG [29]	3	2.516	3.037	5.094
+ LG [29]	5	2.553	3.052	5.173
CA-DINO				
+ DLG	动态	2.186	2.473	4.306
++ LC	动态	**2.021**	**2.331**	**4.012**

基于阶段 I 中由预训练模型生成的伪标签，我们开展了一些实验，以说明所提出方法的有效性。相应结果如表VII所示。首先，遵循 [25],提出的迭代学习框架，我们根据 CA-DINO 提取的说话人嵌入估计伪标签，并使用这些标签训练新的编码器。

表VIII：不同迭代次数下，所提出的 DLG-LC 与其他策略在 Vox-O、Vox-E 和 Vox-H 测试集上的等错误率（%）和最小检测代价函数（p=0.01）比较。未使用 DLC-LC 的 SimCLR 和 CA-DINO 表示我们在训练过程中使用了数据的全部估计伪标签，未进行数据筛选。

初始模型	DLG-LC	迭代	Vox-O		Vox-E		Vox-H	
			等错误率(%)	最小检测代价函数	等错误率(%)	最小检测代价函数	等错误率(%)	最小检测代价函数
SimCLR	×	初始	8.547	0.6453	9.228	0.6912	14.21	0.7757
		1	6.281	0.5811	7.428	0.6221	11.54	0.7213
		2	5.914	0.5299	6.745	0.5880	10.54	0.6971
		3	5.547	0.5259	6.407	0.5580	10.14	0.6698
		4	4.872	0.4651	5.593	0.5144	8.923	0.6408
		5	**4.484**	**0.4545**	**5.225**	**0.5055**	**8.501**	**0.6321**
CA-DINO	×	初始	3.585	0.3529	3.852	0.4182	6.918	0.5743
		1	2.909	0.3000	3.315	0.3372	5.692	0.4654
		2	2.606	0.2887	3.181	0.3211	5.403	0.4489
		3	**2.558**	0.3054	**3.064**	**0.3176**	5.342	**0.4482**
		4	2.643	**0.2825**	3.065	0.3200	**5.291**	0.4483
CA-DINO	✓	初始	3.585	0.3529	3.852	0.4182	6.918	0.5743
		1	2.021	0.2171	2.331	0.2419	4.012	0.3484
		2	1.596	0.1665	2.004	0.2089	3.484	0.3083
		3	**1.585**	0.1671	**1.879**	**0.1963**	3.293	**0.2941**
		4	1.606	**0.1636**	1.906	0.2028	**3.274**	0.2955

为了体现我们方法的优越性，我们还基于 SimCLR 训练了一个模型。SimCLR 是最流行的自监督说话人验证方法 [21]。从表中的结果可以看出，基于 CA-DINO 的模型在所有测试集上都以大幅提升超过了 SimCLR。随后，基于预训练的 CA-DINO，我们还在迭代 1 中对 DLG-LC 进行了探索。根据结果可以观察到，与未进行任何数据选择训练的系统相比，采用固定阈值进行数据选择的损失门控（LG）学习能够带来显著提升。这意味着损失门控能够有效选择有利于模型的可靠标签。然而，我们还尝试设置不同的阈值（1、3、5），发现阈值的选择也会对模型性能产生不可忽视的影响 [29]。基于估计得到的高斯混合模型，我们提出的动态损失门控（DLG）能够结合当前训练情况动态调整阈值，其性能优于在整个训练过程中仅采用固定阈值的 LG。此外，我们应用标签校正（LC）策略，充分利用带有不可靠标签的数据，结果进一步得到提升。与基线系统（不进行数据选择的 SimCLR）相比，所提出的采用 DLG-LC 的 CA-DINO 在 Vox-O、Vox-E 和 Vox-H 测试集上的等错误率分别相对降低了 **70.05%**、**68.61%**和**65.23%** 。

D. 使用 *DLG-LC* 进行迭代学习

为了进一步说明我们所提方法的优越性，在 [25]之后，我们进行了多轮迭代训练。我们汇总了在 Vox-O、Vox-E、Vox-H 测试集上，采用或不采用所提出的 DLG- LC 策略时每次迭代的等错误率和最小检测代价函数性能，结果如表VIII所示。首先，我们分别比较了 SimCLR 和 CA-DINO 的迭代结果，值得注意的是，两者的训练均未采用任何损失门控策略。

根据结果可以观察到，随着迭代次数的增加，迭代学习方法能够持续提升系统性能。然而，基于 SimCLR 的收敛速度明显慢于基于 CA-DINO 的收敛速度。SimCLR 即使在第 5 轮时也尚未收敛，而 CA-DINO 在第 3 轮时已达到最佳性能。此外，采用迭代学习的 SimCLR 的最终性能甚至仍然劣于 CA-DINO 的初始性能。在每次迭代中，所提出的 CA-DINO 始终显著优于 SimCLR，这进一步证明了所提出的 CA-DINO 在自监督说话人验证中的优越性。

基于 CA-DINO 生成的伪标签，我们应用了所提出的 DLG-LC 策略，性能进一步显著提升。仅需一轮迭代即可获得优于不使用 DLC-LC 时三轮迭代的结果，这表明动态阈值筛选和标签校正对于数据利用的重要性。在经过更多迭代并收敛后，其性能远胜于不使用DLG-LC的系统。这表明，所提出的DLG-LC不仅能够加速模型收敛、缩短训练时间，还能显著提升自监督学习模型的性能上限。

E. 融入多模态

然后，我们在迭代学习过程中引入视觉信息。与 [30] 中的工作不同，我们不仅在进行数据聚类时使用多模态，还通过 DLG-LC 在应用数据选择时利用多模态信息。表IX展示了 DLG-LC 在单模态和多模态下的等错误率和最小检测代价函数性能对比。

观察发现，在迭代学习中融入音频和视觉模态知识，可进一步大幅提升性能，这表明额外的视觉信息

表 IX：采用单模态或多模态的所提出 DLG-LC 在不同迭代次数下，于 Vox-O、Vox-E、Vox-H 测试集上的等错误率（%）和最小检测代价函数（p=0.01）对比。需要指出的是，它们在第一个自监督预训练阶段均以 CA-DINO 初始化。我们的音频和视觉编码器均独立训练，只有在迭代学习过程中进行聚类数据和数据选择时，才会融合多模态信息。测试时仍仅使用单一音频模态。

训练模态	迭代	Vox-O		Vox-E		Vox-H	
		等错误率(%)	最小检测代价函数	等错误率（%）	最小检测代价函数	等错误率(%)	最小检测代价函数
音频	初始	3.585	0.3529	3.852	0.4182	6.918	0.5743
音频	1	2.021	0.2171	2.331	0.2419	4.012	0.3484
	2	1.596	0.1665	2.004	0.2089	3.484	0.3083
	3	**1.585**	0.1671	**1.879**	**0.1963**	3.293	**0.2941**
	4	1.606	**0.1636**	1.906	0.2028	**3.274**	0.2955
音视频	1	1.537	**0.1326**	1.789	0.1910	3.235	0.3007
	2	**1.292**	0.1565	**1.571**	**0.1688**	**2.799**	**0.2676**
	3	1.356	0.1553	1.602	0.1711	2.839	0.2712

表 X：所提出的结合 DLG-LC 的 CA-DINO 与其他最先进的自监督系统在 Vox-O、Vox-E 和 Vox-H 上的等错误率（%）对比。文中详细列出了各系统的模型架构、聚类数量、方法和迭代轮数。请注意，此处的 AHC 和 K-M 分别指层次凝聚聚类和 k- 均值。此处的 ECAPA-S （小型）和 ECAPA-L （大型）分别表示具有 512 个信道和 1024 个信道的 ECAPA-TDNN。

方法	模型	# 迭代	# 簇	聚类	Vox-O（等错误率）	Vox-E（等错误率）	Vox-H（等错误率）
全监督 [53]	ECAPA-S	-	-	-	1.010	1.240	2.320
IDLab [26]	ECAPA-L	7	7500	AHC	2.100	-	-
JHU [27]	Res2Net50	5	7500	AHC	1.890	-	-
SNU [28]	ECAPA-L	5	7500	AHC	1.660	-	-
LG [29]	ECAPA-L	5	6000	K-M	1.660	2.180	3.760
DKU + 单模态 [30]	ResNet34	5	6000	K-M	2.740	3.080	5.480
DKU + 多模态 [30]	ResNet34	5	6000	K-M	1.920	2.030	3.720
CA-DINO	ECAPA-S	3	7500	K-M	2.558	2.129	5.148
CA-DINO +DLG-LC +单模态	ECAPA-S	3	7500	K-M	1.585	1.879	3.293
CA-DINO +DLG-LC +多模态	ECAPA-S	2	7500	K-M	**1.292**	**1.571**	**2.799**
CA-DINO + DLG-LC + 多模态*	ECAPA-S	2	7500	K-M	**1.191**	**1.474**	**2.543**

* 为了与全监督系统 [53]进行公平比较，结果采用自适应 s-norm [60] 给出。

能够提高数据利用效率。以 Vox-H 的等错误率为例，仅使用单一模态的音频数据时，前两次迭代以及第三次迭代的相对等错误率降幅，在 Vox-H 测试集上分别为 **42.01%**、**13.16%** 和 **5.48%** 。若采用音频-视觉数据进行迭代学习，前两次迭代的相对等错误率降幅分别为 **53.24%** 和 **13.48%** 。

F. 与其他系统的比较

本节在表 X 中对我们提出的CA-DINO与DLG-LC及其他自监督说话人验证系统进行了性能比较，其中大多数系统来自最新的VoxCeleb 说话人识别挑战赛（VoxSRC）[61]，[62] ，这些系统代表了如今最先进的系统。此外，全监督系统也作为表X 的第一行列出，以供比较。

与采用大规模模型的先前工作相比，我们采用的模型是 ECAPA-S（Small，C=512），其参数更少，所需的计算资源也更少。与 AHC（层次凝聚聚类）相比，为了更便于实现，我们采用一种更简单、更方便的聚类方法 K-M（k-均值）来生成伪标签。此外，在进行聚类数据处理时，我们将簇的数量设置为7500而不是6000，因为6000接近训练集中的真实说话人数量（5994），这一设置过于特殊。从结果来看，可以明显观察到，我们提出的新型自监督说话人验证框架在单模态和多模态方面都远远优于所有现有方法，即使迭代次数更少、模型更小且聚类方法更简单。在单模态条件下，所提出的采用DLG-LC的CA-DINO仅经过3次迭代，就在Vox-O、Vox-E和Vox-H数据集上分别以相对提升 [29] 超过最佳系统（LG）**4.52%**、**13.81%** 和**12.42%** 。如果我们在迭代学习阶段使用音视频数据，相应的提升将扩大到相对**22.17%**、**27.94%** 和**25.56%**，这是一次巨大的性能飞跃。

总之，尽管我们使用更少的迭代次数、更小的模型和更简单的聚类方法来训练系统，所提出的系统仍以大幅提升的性能，在自监督说话人验证方面达到了新的**最先进水平**。更值得期待的是，与采用 ECAPA-TDNN-Small 的传统全监督系统相比，我们新提出的自监督学习系统甚至取得了可与其相媲美的性能，但

无需使用任何真实标签。

VII. 结论

本文提出了一种先进的簇感知 DINO（CA-DINO），并将动态损失门控与标签校正（DLG-LC）用于自监督说话人验证。引入 DINO 框架后，系统无需负样本即可进行训练，与其他自监督模型相比有了显著改进。随后，我们将聚类感知训练融入 DINO 框架，并从同一类别中收集正样本，而非仅使用单个句子中的样本，使模型能够利用更多样化的数据并提升系统性能。在迭代学习阶段，我们通过用高斯分布对损失直方图进行建模，得到动态损失门控，以便在使用伪标签训练时筛选可靠数据。我们不直接丢弃不可靠数据，而是采用预测后验作为目标分布，以避免模型拟合错误样本。此外，我们将多模态信息融入 DLG-LC，进一步提升性能。在 VoxCeleb 上进行的实验表明，我们新提出的采用 DLG-LC 的 CA-DINO 表现优异，在自监督说话人验证方面达到了新的**最先进水平**。更值得期待的是，说话人验证中无监督与有监督表征学习之间的差距大幅缩小，并且采用我们的说话人验证自监督学习方法，取得了接近全监督系统的性能。

致谢

本工作部分得到了中国国家自然科学基金项目（批准号：62122050 和 62071288）的资助，部分得到了上海市科技重大项目（批准号：2021SHZDZX0102）的资助。实验在上海交通大学的 PI超级计算机上完成。

参考文献

[1] B. Han、Z. Chen 和 Y. Qian，“使用动态损失门控和标签校正的自监督说话人验证”，发表于*ISCA Interspeech*论文集，2022年。[2] E. Variani、X. Lei、E. McDermott、I. L. Moreno 和 J. Gonzalez-Dominguez，“用于小型文本相关说话人验证的深度神经网络”，发表于*IEEE ICASSP*论文集，2014年，第4052–4056页。[3] D. Snyder、D. Garcia-Romero、G. Sell、D. Povey 和 S. Khudanpur，“X向量：用于说话人识别的鲁棒DNN嵌入”，发表于*IEEE ICASSP*论文集。IEEE，2018年，第5329–5333页。[4] H. Zeinali、S. Wang、A. Silnova、P. Matˇejka 和 O. Plchot，“BUT系统在 VoxCeleb 说话人识别挑战赛 2019中的系统描述”，*arXiv*预印本 *arXiv:1910.12592*，2019年。[5] B. Han、Z. Chen 和 Y. Qian，“使用自注意力进行说话人验证的局部信息建模”，发表于*IEEE ICASSP*论文集。IEEE，2022年，第6727–6731页。[6] B. Han、Z. Chen、B. Liu 和 Y. Qian，“MLP-SVNet：一种基于多层感知机的说话人验证网络”，发表于 *IEEE ICASSP*论文集。IEEE，2022年，第7522–7526页。[7] Y. Liu、L. He 和 J. Liu，“用于说话人验证的大间隔Softmax损失”，*arXiv*预印本 *arXiv:1904.03479*，2019年。[8] J. S. Chung、J. Huh、S. Mun、M. Lee、H. S. Heo、S. Choe、C. Ham、S. Jung、B.-J. Lee 和 I. Han，“为说话人识别中的度量学习辩护”，*arXiv*预印本 *arXiv:2003.11982*，2020年。[9] G. Heigold、I. Moreno、S. Bengio 和 N. Shazeer，“端到端文本相关说话人验证”，发表于*IEEE ICASSP*论文集。IEEE，2016年，第5115–5119页。[10] S. Wang、Y. Yang、Y. Qian 和 K. Yu，“重新审视深度说话人嵌入学习中的统计池化层”，发表于2021年第12届国际中文口语处理研讨会（ISCSLP）。IEEE，2021年，第1–5页。[11] Y. Zhu、T. Ko、D. Snyder、B. Mak 和 D. Povey，“用于文本无关说话人验证的自注意力说话人嵌入”，发表于ISCA Interspeech论文集，第2018卷，2018年，第3573–3577页。[12] D. A. Reynolds、T. F. Quatieri 和 R. B. Dunn，“使用自适应高斯混合模型进行说话人验证”，《数字信号处理》，第10卷，第1–3期，第19–41页，2000年。[13] N. Dehak、P. Kenny、R. Dehak、P. Dumouchel 和 P. Ouellet，“用于说话人验证的前端因子分析”，《IEEE/ACM音频、语音与语言处理汇刊》，第19卷，第4期，第788–798页，2011年。[14] A. Baevski、Y. Zhou、A. Mohamed 和 M. Auli，“wav2vec 2.0：语音表征自监督学习框架”，《神经信息处理系统大会论文集》，第33卷，第12 449–12 460页，2020年。[15] W.-N. Hsu、B. Bolte、Y.-H. H. Tsai、K. Lakhotia、R. Salakhutdinov 和 A. Mohamed，“HuBERT：通过隐藏单元掩码预测进行自监督语音表征学习”，《IEEE/ACM音频、语音与语言处理汇刊》，第29卷，第3451–3460页，2021年。[16] Z. Fan、M. Li、S. Zhou 和 B. Xu，“探索 wav2vec 2.0在说话人验证和语言识别中的应用”，arXiv预印本 arXiv:2012.06185，2020年。[17] Z. Chen、S. Chen、Y. Wu、Y. Qian、C. Wang、S. Liu、Y. Qian 和 M. Zeng，“用于自动说话人验证的大规模自监督语音表征学习”，发表于IEEE ICASSP论文集。IEEE，2022年，第6147–6151页。[18] T. Stafylakis、J. Rohdin、O. Plchot、P. Mizera 和 L. Burget，“自监督说话人嵌入”，arXiv预印本 arXiv:1904 03486，2019年。.[19] N. Inoue 和 K. Goto，“ 具有广义对比损失的半监督对比学习及其在说话人识别中的应用”，发表于IEEE APSIPA ASC论文集。IEEE，2020年，第1641–1646页。[20] J. Huh、H. S. Heo、J. Kang、S. Watanabe 和 J. S. Chung，“用于无监督说话人识别的增强对抗训练”，arXiv预印本 arXiv:2007.12085，2020年。[21] H. Zhang、Y. Zou 和 H. Wang，“用于文本无关说话人验证的对比式自监督学习”，发表于IEEE ICASSP论文集。IEEE，2021年，第6713–6717页。[22] W. Xia、C. Zhang、C. Weng、M. Yu 和 D. Yu，“使用原型动量对比学习的自监督文本无关说话人验证”，发表于IEEE ICASSP论文集。IEEE，2021年，第6723–6727页。[23] S. H. Mun、W. H. Kang、M. H. Han 和 N. S. Kim，“通过对比均衡学习进行说话人识别的无监督表征学习”，arXiv预印本 arXiv:2010.11433，2020年。[24] M. Caron、P. Bojanowski、A. Joulin 和 M. Douze，“用于视觉特征无监督学习的深度聚类”，发表于ECCV论文集，2018年，第132–149页。[25] D. Cai、W. Wang 和 M. Li，“一种用于自监督深度说话人表征学习的迭代框架”，发表于IEEE ICASSP论文集。IEEE，2021年，第6728–6732页。[26] J. Thienpondt、B. Desplanques 和 K. Demuynck，“IDLab VoxCeleb 说话人识别挑战赛 2020系统描述”，arXiv预印本 arXiv:2010.12468，2020年。[27] J. Cho、J. Villalba 和 N. Dehak，“JHU提交的VoxSRC-21：赛道3”，arXiv预印本 arXiv:2109.13425，2021年。[28] S. H. Mun、M. H. Han 和 N. S. Kim，“SNU-HIL在VoxCeleb 说话人识别挑战赛 2021中的系统”，VoxSRC，2021年。[29] R. Tao、K. A. Lee、R. K. Das、V. Hautam¨aki 和 H. Li，“基于损失门控学习的自监督说话人识别”，arXiv预印本 arXiv:2110.03869，2021年。[30] D. Cai、W. Wang 和 M. Li，“将视觉信息融入基于音频的自监督说话人识别”，《IEEE/ACM音频、语音与语言处理汇刊》，第30卷，第1422–1435页，2022年。[31] M. Caron、H. Touvron、I. Misra、H. J´egou、J. Mairal、P. Bojanowski 和 A. Joulin，“自监督视觉Transformer中的新兴属性”，发表于ICCV论文集，2021年，第9650–9660页。[32] D. Berthelot、N. Carlini、E. D. Cubuk、A. Kurakin、K. Sohn、H. Zhang 和 C. Raffel，“RemixMatch：通过分布对齐和增强锚定实现半监督学习”，arXiv预印本 arXiv:1911.09785，2019年。[33] K. Sohn、D. Berthelot、N. Carlini、Z. Zhang、H. Zhang、C. A. Raffel、E. D. Cubuk、A. Kurakin 和 C.-L. Li，“FixMatch：通过一致性和置信度简化半监督学习”，《神经信息处理系统大会论文集》，第33卷

，第596–608页，2020.

[34] B. Zhang, Y. Wang, W. Hou, H. Wu, J. Wang, M. Okumura 和 T. Shi- nozaki，“Flexmatch：通过课程式伪标签提升半监督学习”，《神经信息处理系统大会论文集》，第 34 卷，2021 年。[35] A. Nagrani, J. S. Chung, S. Albanie 和 A. Zisserman，“使用跨模态自监督的解耦语音嵌入”，载于《IEEE ICASSP 论文集》。IEEE，2020 年，第 6829–6833 页。[36] S.-W. Chung, H. G. Kang 和 J. S. Chung，“看见声音与听见声音：使用跨模态自监督学习判别性嵌入”，arXiv 预印本 arXiv:2004.14326，2020 年。[37] T. Chen, S. Kornblith, M. Norouzi 和 G. Hinton，“视觉表征对比学习的简单框架”，载于《国际机器学习会议论文集》。PMLR，2020 年，第 1597–1607 页。[38] K. He, H. Fan, Y. Wu, S. Xie 和 R. Girshick，“用于无监督视觉表征学习的动量对比”，载于《计算机视觉与模式识别会议论文集》，2020 年 6 月。[39] D. Cai 和 M. Li，“用于 2021 年 VoxCeleb 说话人识别挑战赛自监督说话人验证任务的 DKU-DukeECE 系统”，arXiv 预印本 arXiv:2109.02853，2021 年。[40] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski 和 A. Joulin，“通过对比群组分配进行无监督视觉特征学习”，载于《神经信息处理系统大会论文集》，2020 年。[41] G. Hinton, O. Vinyals, J. Dean 等，“提炼神经网络中的知识”，arXiv 预印本 arXiv:1503.02531，第 2 卷，第 7 期，2015 年。[42] I. Loshchilov 和 F. Hutter，“SGDR：带热重启的随机梯度下降”，arXiv 预印本 arXiv:1608.03983，2016 年。[43] J. S. Chung, A. Nagrani 和 A. Zisserman，“VoxCeleb2：深度说话人识别”，载于《ISCA Interspeech 论文集》，2018 年，第 1086–1090 页。[44] E. Arazo, D. Ortego, P. Albert, N. O’Connor 和 K. McGuinness，“无监督标签噪声建模与损失校正”，载于《国际机器学习会议论文集》。PMLR，2019 年，第 312–321 页。[45] X. Xiang, S. Wang, H. Huang, Y. Qian 和 K. Yu，“间隔至关重要：面向说话人识别的更具判别性的深度神经网络嵌入”，载于《IEEE APSIPA ASC 论文集》。IEEE，2019 年，第 1652–1656 页。[46] F. Tong, Y. Liu, S. Li, J. Wang, L. Li 和 Q. Hong，“使用噪声标签进行说话人嵌入学习的自动纠错”，载于《ISCA Interspeech 论文集》，2021 年，第 4628–4632 页。[47] D.-H. Lee 等，“伪标签：用于深度神经网络的简单高效半监督学习方法”，载于《国际机器学习会议论文集》，第 3 卷，2，2013 年，896 no. p.。[48] A. Nagrani, J. S. Chung 和 A. Zisserman，“VoxCeleb：大规模说话人识别数据集”，载于《ISCA Interspeech 论文集》，2017 年，第 2616–2620 页。[49] W. Cai, J. Chen, J. Zhang 和 M. Li，“面向说话人和语言识别的即时数据加载器与话语级聚合”，《IEEE/ACM 音频、语音与语言处理汇刊》，第 28 卷，第 1038–1051 页，2020 年。[50] D. Snyder, G. Chen 和 D. Povey，“MUSAN：音乐、语音和噪声语料库”，arXiv 预印本 arXiv:1510.08484，2015 年。[51] T. Ko, V. Peddinti, D. Povey, M. L. Seltzer 和 S. Khudanpur，“面向鲁棒语音识别的混响语音数据增强研究”，载于《IEEE ICASSP 论文集》。IEEE，2017 年，第 5220–5224 页。[52] K. Zhang, Z. Zhang, Z. Li 和 Y. Qiao，“使用多任务级联卷积网络进行联合人脸检测与对齐”，《IEEE 信号处理快报》，第 23 卷，第 10 期，第 1499–1503 页，2016 年。[53] B. Desplanques, J. Thienpondt 和 K. Demuynck，“ECAPA-TDNN：基于时延神经网络的强化通道注意力、传播与聚合，用于说话人验证”，arXiv 预印本 arXiv:2005.07143，2020 年。[54] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang 和 P. H. Torr，“Res2Net：一种新的多尺度骨干架构”，《IEEE 模式分析与机器智能汇刊》，2019 年。[55] J. Hu, L. Shen 和 G. Sun，“挤压-激励网络”，载于《计算机视觉与模式识别会议论文集》，2018 年，第 7132–7141 页。[56] J. Johnson, M. Douze 和 H. J´egou，“使用图形处理器进行十亿规模的相似性搜索”，《IEEE 大数据汇刊》，第 7 卷，第 3 期，第 535– 547 页，2019 年。[57] K. He, X. Zhang, S. Ren 和 J. 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。[58] Y. Kim, W. Park, M.-C. Roh 和 J. Shin，“GroupFace：学习潜在群组并构建用于人脸识别的基于群组的表征”，载于《计算机视觉与模式识别会议论文集》，2020 年，第 5621–5630 页。[59] J. Deng, J. Guo, N. Xue 和 S. Zafeiriou，“ArcFace：用于深度人脸识别的加性角度间隔损失”，载于《计算机视觉与模式识别会议论文集》，2019 年，第 4690– 4699 页。[60] P. Mat˘ejka, O. Novotn´y, O. Plchot, L. Burget, M. D. S´anchez 和 J. ˇCernock´y，“多语言说话人识别中的分数归一化分析”，载于《ISCA Interspeech 论文集》，2017 年，第 1567–1571 页。

[61]A. Nagrani、J. S. Chung、J. Huh、A. Brown、E. Coto、W. Xie、M. McLaren、D. A. Reynolds 和 A. Zisserman，“Voxsrc 2020：第二届 VoxCeleb 说话人识别挑战赛”，*arXiv* 预印本 *arXiv:2012.06867*，2020。[62] A. Brown、J. Huh、J. S. Chung、A. Nagrani 和 A. Zisserman，“Voxsrc 2021：第三届 VoxCeleb 说话人识别挑战赛”，*arXiv* 预印本 *arXiv:2201.04583*，2022。