arXiv:2302.02845v1 [cs.SD] 6 Feb 2023

通过将视频蒸馏为特权信息进行音频表示学习

Amirhossein Hajavi，*IEEE* 学生会员，Ali Etemad，*IEEE* 高级会员

摘要—**与单模态方法相比，使用多模态视听数据进行深度音频表示学习通常能取得更好的性能。然而，在现实场景中，推理时并非总能获取这两种模态，这会导致为多模态推理训练的模型性能下降。在本研究中，我们提出一种新颖的深度音频表示学习方法，在推理时视频模态缺失的情况下使用视听数据。为此，我们采用特权信息学习（LUPI）框架下的教师-学生知识蒸馏。以往针对 LUPI 提出的方法使用教师生成的软标签，而我们的方法则使用**教师学习到的嵌入**来训练学生网络。我们将该方法应用于两种不同的设置：一种是序列数据，特征在时间维度上划分为多个片段；另一种是非序列数据，将全部特征视为一个整体片段。在非序列设置中，教师网络和学生网络均由编码器组件和任务头组成。我们使用教师编码器组件生成的嵌入来训练学生编码器，而学生的任务头则使用真实标签进行训练。在序列设置中，网络还增加了一个位于编码器与任务头之间的聚合组件。我们使用教师的编码器和聚合组件生成的两组嵌入来训练学生。与非序列设置类似，学生网络的任务头使用真实标签进行训练。我们在两种不同的视听任务上测试了该框架，即说话人识别和语音情感识别。实验表明，将视频模态视为音频表征学习这一主要目标的特权信息后，与仅基于音频的识别以及以往使用 LUPI 的工作相比，我们的方法取得了显著提升。**

索引词—**深度学习、利用特权信息进行学习、知识蒸馏、多模态数据、视听表示学习。**

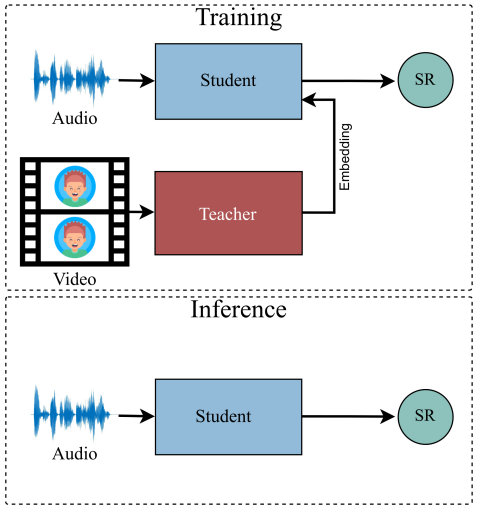

图 1. 所提方法概览。教师从视频模态中提取的嵌入在训练学生时作为特权信息，以提升其学习音频表征的能力。在推理阶段，只有音频模态可用，学生模型负责生成音频嵌入，随后用于执行 SR 和 SER。

一、引言

DEEP 音频表征学习近来备受关注，尤其是在说话人识别（SR） [9], [10], [43], [31], [11] 和语音情感识别（SER） [1], [18], [25]等应用中。深度音频表示学习旨在从音频或视觉信号中学习嵌入，以用于检索说话人的身份或情绪状态等信息。通常，使用多模态音视频输入能最好地实现这一目标 [31], [19], [2] ，而非仅使用单一模态的音频或视频[9], [10], [43], [23], [24], [11]。然而，在许多现实场景中，推理阶段可能无法同时获得这两种模态，导致模型无法有效运行。为解决这一问题，我们提出如下问题：“如何有效地利用两种模态进行训练，从而使单一模态推理受益？”

Vapnik 等人 [42] 将仅在训练期间（而非推理时）可用的信息定义为“特权信息”。他们提出了一种名为“利用特权信息学习”（LUPI）的新学习范式：通过优化松弛变量来降低问题的复杂度，并利用基于特权信息训练的辅助 SVM 模型，帮助主 SVM 更好地完成任务。后来，这一范式被应用于深度神经网络 [28],，研究表明 LUPI 可以通过 Hinton 等人 [15] 提出的知识蒸馏技术实现。在他们的研究中，教师模型利用文本模态中的特权信息进行训练。随后，他们将教师模型和真实标签一起用于训练学生模型，以执行图像分类。

在本研究中，为了实现单模态音频表征

作者感谢 IMRSV Data Labs 对本研究的支持。作者还感谢加拿大自然科学与工程研究委员会（NSERC）对本研究的资助（项目编号：CRDPJ 533919-18）。

本文作者就职于加拿大金斯顿女王大学电气与计算机工程系。（电子邮箱： {a.hajavi, ali.etemad}@queensu.ca）。

学习，同时利用多模态视听流进行训练，我们提出了一种创新方案：采用特权信息，并将视频视为特权信息。我们的方法概览如图1所示。我们的模型基于教师-学生知识蒸馏构建，使视频流能够在训练期间与主要的音频模态一同学习。首先，我们将视频流用作特权信息来训练教师网络。接下来，我们以音频作为输入，同时将真实输出标签和从教师模型获得的视频嵌入作为输出，训练学生网络。通过这种方式，学生模型在推理时仅依赖音频信号即可运行，同时在训练期间利用音频和视频两种模态进行训练并从中受益。我们使用多个广泛应用的视听数据集开展了大量实验，以证明所提出的框架在利用特权信息进行音频表征学习方面的有效性。

总而言之，我们做出了以下贡献：

- 我们提出了一种针对语音识别和语音情感识别的深度音频表示学习新方案，在网络训练过程中利用视频作为特权信息，以提升性能。
- 我们在模型上进行了语音识别和语音情感识别实验，发现与单模态基线相比，性能有显著提升。我们还将本方法与其他基于特权信息的研究进行了比较，结果表明，本方法在音频表征学习方面表现更优。
- 我们通过调整特权信息对网络训练的影响程度，分析了其影响。

本文其余部分安排如下。第 II 节介绍特权信息和知识蒸馏方面的相关工作。第 III 节详细介绍我们提出的方案。随后，我们详细阐述所进行的实验，并在第 IV 节报告结果。最后，我们总结本文并探讨未来可能的研究方向。

II. 相关工作

文献中采用了不同的方法，借助特权信息进行学习 [1], [5], [6], [7], [33], [35], [41], [29], [40], [30], [32], [37], [26]。我们可以将这些研究分为两大类：（1）通过教师-学生模型，依赖知识蒸馏技术进行深度表征学习的研究；（2）利用特权信息但不采用知识蒸馏的研究。鉴于知识蒸馏与特权信息学习（LUPI）及我们的方法密切相关，下面我们首先回顾知识蒸馏的一般概念，随后回顾采用和未采用知识蒸馏的 LUPI 研究。虽然我们的工作属于音频表征学习领域，但鉴于利用特权信息进行音频学习的相关研究数量较少，我们将讨论范围扩展到其他模态，以提供更全面的综述。

A. 知识蒸馏

Hinton 等人 [15]提出了知识蒸馏，使较小的机器学习模型（称为“学生”）能够向较大的机器学习模型（称为“教师”）学习。此后，大量研究探索了知识蒸馏在深度表征学习中的应用。该领域的开创性工作包括 [36], [17], [20], [47], [13], [44], [8], [14]。这些研究表明，知识蒸馏能够使学生模型达到与教师模型相当的性能，同时减少可学习参数的数量，从而降低计算负担和所需内存。

知识蒸馏通常通过两种主要方法进行。第一种方法是使用从教师模型获得的“软标签”来训练学生 [45], [46],；而第二种方法则依靠教师模型学习到的“嵌入”来训练学生 [36], [17], [20], [47], [13], [44], [8]。除了这两种方法，文献中还提出了一些其他方案。例如， [14]将教师模型中神经元的激活边界迁移给学生，以缩短训练时间。在 [22], [34], 中，研究者通过将教师模型生成的注意力图迁移给学生来进行知识蒸馏。这种技术借助迁移得到的注意力图，帮助学生找到输入中的显著区域，从而提升学生的性能，而无需额外训练。在 [4], 中，研究者不再仅依赖单个嵌入，而是利用教师模型多个层中的知识来训练学生，从而提升学生在推理过程中的泛化能力。

B. 基于知识蒸馏的特权信息学习

作为最早将特权信息学习与深度神经网络结合的解决方案之一， [28]提出了采用教师-学生知识蒸馏的方法。他们提出，教师模型可以接收来自特权信息的输入，并将其输出与真实标签结合起来，训练接收主要数据作为输入的学生模型。近期的许多研究也采用了类似技术 [1], [5], [6], [7], [33], [35], [41], [29], [40], [30], [32]。这些方法的核心思想是，将次要模态作为教师模型的输入，并将其用作特权信息的来源。

此前只有极少数研究在考虑将「视频」作为次要模态的情况下，针对音频表征学习展开研究 [1], [30], [32]。在这些方法中，教师模型将视频帧作为输入，并生成软标签，随后软标签被用作训练学生模型的唯一来源。这些研究的主要目标是通过仅使用由次要模态（视频）生成的软标签来训练学生模型，从而缓解对主要模态（音频）带标签训练数据的需求。尽管这些研究成功地实现了使用视频作为 [自 -] 监督信号来训练音频的目标，但它们并未探索将次要模态用作特权信息、以帮助网络学习主要模态的影响。然而，在本研究中，我们旨在利用次要模态，

在使用主要模态的标注数据的同时提升网络性能，而不是完全避免使用输出标签。

C. 不使用知识蒸馏的特权信息学习

此外，许多研究也采用教师-学生框架以外的方法，利用特权信息训练网络。在 [37], 所开展的研究中，采用了多任务学习来训练模型。其所提网络在执行动作识别的同时，还旨在从另一种模态重建特权信息。该研究使用的主要模态是个体执行特定动作的视频帧；作为特权信息，则使用视频中个体的骨骼关节位置。这项工作中对特权信息的使用，与若干基线方法相比提升了性能。虽然所提方法已证明对维度有限的次要模态有益，但尚未探索将其与维度较高的次要模态（例如视频帧）结合。在 [26], 该研究使用由特权信息生成的丢弃掩码，以提升深度神经网络的泛化能力。研究使用从输入中获取的图像片段作为特权信息，生成启发式丢弃掩码。随后，这些掩码被应用于深度神经网络学习到的表示，以帮助模型实现泛化。尽管该方法已被证明能够提升模型的泛化能力，但丢弃掩码是利用原始输入的一个片段生成的。因此，该方法要求特权信息与原始输入来自相同模态。

III. 方法

本研究的目标是训练一个深度神经网络，利用视听数据学习音频表征，同时确保在推理/测试期间视频模态不可用。我们的方法采用教师-学生知识蒸馏和 LUPI 范式，训练能够满足这一条件的网络。在我们的方法中，学生模型以音频模态为输入，并利用两种输出进行训练：一种来自真实标签，另一种来自在次要模态上运行的教师模型所获得的嵌入。本节将介绍所提方法的各个组成部分。

A. 预备知识

用于监督学习的训练数据通常以元组的形式呈现（x_i, y_i）$^{i\in\{0,..,n\}}$，其中 x_i 是训练样本的向量表示，y_i 是输出标签。模型的目标是找到一个最优网络$\mathcal{F}$，以尽可能低的误差预测测试数据 $y_i = \mathcal{F}(x_i)$的标签。在这种训练方式中，模型的训练、测试和部署阶段所用的数据格式保持不变。有时，在训练阶段，我们可能会获得测试集中没有的额外信息。Vapnik 等人将这类信息称为特权信息，并提出了一种名为“特权信息学习”的技术，以利用这些信息更好地训练机器学习模型 [42]。通过这种范式训练模型时，会使用元组 (x_i, x_i^*, y_i)，其中 x_i^*表示特权信息。然而，使用训练好的模型进行推理时，数据仍保持原有格式，且无法获取额外信息。

LUPI 范式最初是在支持向量机（SVM）的背景下提出的 [42]。研究表明，该模型即 SVM+ 的表现优于经典 SVM，也就是说，优于未使用特权信息训练的 SVM。LUPI 范式也可以通过教师-学生知识蒸馏实现。按照这种范式的观点，原始的知识蒸馏方法 [15]通过将特权信息用作教师模型的输入而得到扩展。给定元组 (x_i, x_i^*, y_i)，首先使用特权信息与标签组成的元组 (x_i^*, y_i) 训练教师模型。接下来，教师使用特权信息，为每个训练样本生成软标签 s_i，即预测输出概率。最后，学生模型的任意层 L 都使用元组 (x_i, y_i) 和 (x_i, s_i) 同时进行训练，最终梯度 ∇_s计算如下：

$$\nabla_s = (1-\alpha)\nabla\frac{\mathcal{L}_i(y_i', y_i)}{\nabla L} + \alpha\nabla\frac{\mathcal{L}_i(y_i', s_i)}{\nabla L}. \quad (1)$$

其中，参数 α 是模仿参数，用于确定学生网络应在多大程度上遵循教师模型。

B. 我们的解决方案

如前所述，我们的目标是开发一个框架，将视频数据蒸馏到音频学习器中，以便在两种模态均可用时提升训练效果，同时在推理时仅使用音频。为此，我们采用教师-学生知识蒸馏：教师网络以视频数据为输入，并利用从其中间层提取的嵌入来辅助训练以音频数据为输入的学生网络。我们的方法概览已在前文的图1中展示。以下小节将介绍我们方法中各个组成部分的细节。值得注意的是，与大多数音频表征学习工作类似，我们在模型中使用从音频信号提取的频谱图。不过，这些频谱图可以采用两种不同的格式：非序列格式和序列格式。非序列表征将整个频谱图视为一张图像，而在序列版本中，频谱图会沿时间轴划分为较小的片段。因此，我们针对这两种方式（非序列和序列）设计了模型，并将在以下小节中分别介绍这两个版本的教师-学生方法。

1) 视频学习： 下面介绍我们的模型中视频表征学习的过程，该过程由教师网络完成。教师网络的训练过程是单独进行的，并且在学生网络训练之前完成。在学生网络的训练过程中，教师网络处于冻结状态，其权重保持不变。

假设训练样本表示为一个元组集合 (x_i, x_i^*, y_i)，其中 x_i 是来自音频模态的 i^{th} 训练样本， x_i^* 是对应的视频模态样本。在非序列方法中，教师网络由一个编码器 $F^{T_{enc}}$ 以及任务头组成，任务头包含一个全连接层（FC），后接 Softmax 激活，记作 $F^{T_{head}}$。这里，$F^{T_{enc}}$ 使用以下方式为来自x^* 的视频的每一帧生成嵌入

$$E_{i_j}^T = F^{T_{enc}}(x_{i_j}^*), \tag{2}$$

其中， j 是视频帧索引。相应的 $F^{T_{head}}$随后会生成模型的预测标签。教师模型的最终输出 O^T 据此计算为

$$O_{i_j}^T = F^{T_{head}}(E_{i_j}^T). \tag{3}$$

最终输出 $O_{i_j}^T$ 是一个向量，其维度等于模型所考虑的类别数。 $O_{i_j}^T$ 中的每个索引都包含计算得到的概率，表示 $x_{i_j}^*$ 属于该索引所代表类别的概率。

通常，在学习音视频表征时，多个视频帧往往对应于跨越一段时间的频谱图。为解决这一问题，先前的研究提出选择单个帧，并将其称为峰值帧[1]，[30]，[32] ，对应于每个给定的输入频谱图。受这一方法启发，我们在解决方案中按如下方式选择峰值帧嵌入：

$$E_{i_{peak}}^T = E_{i_j}^T | Arg\max_j(Arg\max(O_{i_j}^T)), \tag{4}$$

其中，选择其类别计算概率最高的帧。类别通过对经Softmax函数处理的 $O_{i_j}^T$ 执行Argmax来确定。

另一方面，在序列方法中，我们将整个视频用作教师网络的输入。在此版本中，聚合子网络 F_{agg}^T 被添加到教师网络中，并置于编码器和任务头之间。这里， $F^{T_{agg}}$ 使用以下方式根据编码器的输出生成嵌入：

$$E_i^{T_{agg}} = F^{T_{agg}}(\{E_{i_j}^T | 0 \le j < N_i\}), \tag{5}$$

其中 $E_{i_j}^T$ 是根据公式（2 ）得到的编码器嵌入，N_i 是每个视频中的帧数。

采用序列方法时，我们还会将音频的输入 x_i 划分为固定大小的片段 x_{i_k} ，其中 k 是音频片段的索引。如前所述，跨越一段时间的频谱图通常涵盖视频中的多个帧。因此，每个音频片段 x_{i_k} 都会与多个嵌入 $E_{i_j}^T$ 相匹配。在教师网络的序列版本中，每一帧都没有对应其类别的概率分数，因此无法选择峰值帧。为了解决这个问题，我们将嵌入 $E_{i_k}^T$ 定义为该片段中所有嵌入的平均值 Avg:

$$E_{i_k}^T = Avg(\{E_{i_j}^T | r \times k \le j < r \times (k+1)\}), \tag{6}$$

其中， r 是每个音频片段的帧数。计算得到的嵌入 $E_{i_k}^T$ 随后与音频片段 x_{i_k}相匹配。

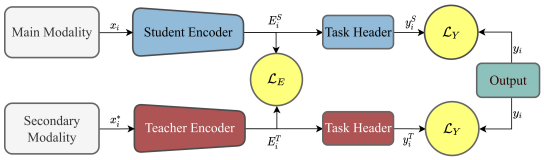

图 2. 在非序列设置中，通过师生蒸馏学习特权信息的所提出框架。在此设置中，完整的音频信号以单个频谱图的形式提供给学生，而教师网络接收视频的单帧作为输入。在这里，学生网络和教师网络均由编码器组件和任务头构成。教师网络的编码器组件生成的嵌入用于部分训练学生网络的编码器组件，所用梯度由 $\mathcal{L}_E$ 生成。同时，学生网络的编码器组件和任务头接收由 $\mathcal{L}_Y$ 生成的梯度。

2）音频学习：在非序列方法中，学生网络将完整的音频信号作为单个频谱图输入。图2 展示了所提方法的非序列版本的总体方案。在此版本中，学生模型由一个编码器、 $F^{S_{enc}}$以及一个任务头组成，该任务头包含一个带 Softmax 激活的全连接层，记为 $F^{S_{head}}$。编码器使用 E_i^S 从输入 x_i 中提取嵌入

$$E_i^S = F^{S_{enc}}(x_i). \tag{7}$$

嵌入 E_i^S 随后被输入 $F^{S_{head}}$ 以预测输出 y_i'。接着，我们定义损失函数 $\mathcal{L}_E(E_i^S, E_{i_{peak}}^T)$，用于计算学生网络编码器组件的嵌入与使用公式计算出的嵌入之间的距离。2。我们定义第二个损失函数$\mathcal{L}_Y(y_i', y_i)$，用于计算 y_i' 与真实标签 y_i之间的距离。学生模型的输出层仅使用 $\mathcal{L}_Y$进行训练，而编码器则通过使用 ∇_s 计算出的梯度进行训练

$$\nabla_s = (1-\alpha)\nabla\frac{\mathcal{L}_Y(y_i', y_i)}{\nabla F^{S_{enc}}} + \alpha\nabla\frac{\mathcal{L}_E(E_i^S, E_{i_{peak}}^T)}{\nabla F^{S_{enc}}}, \tag{8}$$

其中 α 充当模仿参数。 α的值决定了训练学生模型编码器组件时梯度的权重，取值范围为 0 到 1。

在序列方法中，学生模型的音频输入会沿时间维度划分为固定大小的片段。在此，我们在编码器与任务头之间向学生网络添加一个聚合器子网络 $F^{S_{agg}}$ 。编码器 $F^{S_{enc}}$ 使用以下方式从每个音频片段 x_{i_k}生成嵌入 $E_{i_k}^S$

$$E_{i_k}^S = F^{S_{enc}}(x_{i_k}). \tag{9}$$

然后使用以下方式计算最终嵌入 $E_i^{T_{agg}}$

$$E_i^{T_{agg}} = F^{S_{agg}}(\{E_{i_k}^T | 0 \le k < M_i\}), \tag{10}$$

其中 M_i 是 x_i中的音频片段数量。最后，使用以下方式生成学生模型 y_i' 的输出

$$y_i' = F^{S_{head}}(E_i^{T_{agg}}). \tag{11}$$

图 3 展示了训练学生网络的所提方法序列版本的两种可能方案：(1) 在编码器级进行视频信息蒸馏

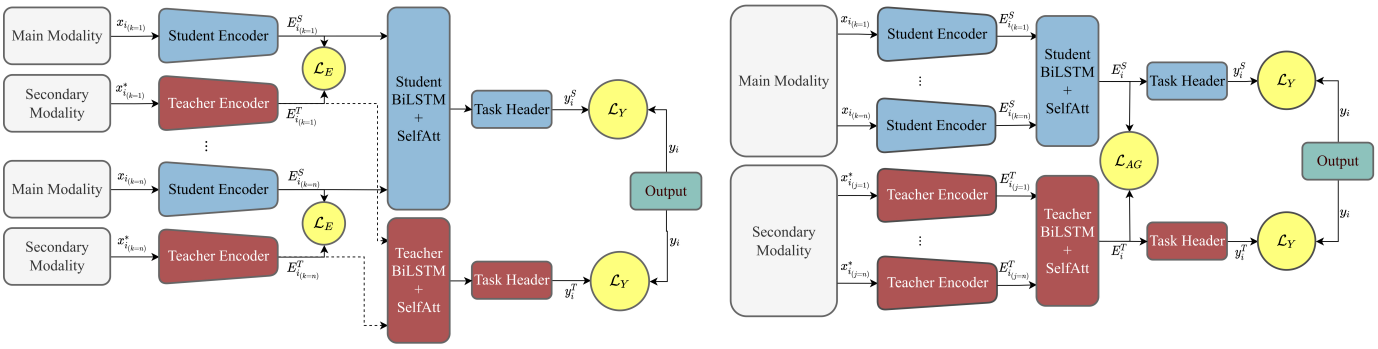

图3. 通过师生蒸馏学习特权信息的所提框架。在我们方法的序列设置中，音频信号沿时间轴划分为多个大小相同的片段。然后，每个片段都会输入学生网络的编码器组件。编码器组件生成的嵌入随后被收集起来，并传递给聚合器，由其从这些嵌入中提取时间依赖关系。另一方面，教师网络会为每个视频帧生成嵌入，并生成整个视频的总体嵌入。在我们方法的编码器级实现（左）中，用于训练学生网络编码器的部分梯度来自 $\mathcal{L}_E$，它会比较学生网络编码器生成的嵌入与教师网络编码器生成的嵌入。在所提方法的聚合器级实现（右）中，学生网络的编码器和聚合器组件都会通过 $\mathcal{L}_{AG}$ 从特权信息中获得梯度。该损失比较学生网络聚合器生成的嵌入与教师网络聚合器生成的嵌入。在这两种实现中，用于学生网络其余部分的另一部分梯度则由真实标签生成。

表I VGG 非序列学生网络的架构详情。

层		激活函数	形状
Input		–	$257 \times 500 \times 1$
$Conv2D + BN$		ReLU	$257 \times 250 \times 64$
$Maxpool$		–	$128 \times 250 \times 64$
$Conv2D + BN$		ReLU	$128 \times 250 \times 128$
$Maxpool$		–	$64 \times 125 \times 128$
$Conv2D + BN$		ReLU	$64 \times 125 \times 128$
$Maxpool$		–	$32 \times 62 \times 128$
$Conv2D + BN$		–	$32 \times 62 \times 256$
$Conv2D$		ReLU	$32 \times 62 \times 256$
$Maxpool$		–	$16 \times 31 \times 256$
$Conv2D + BN$		ReLU	$16 \times 31 \times 512$
$Maxpool$		–	$8 \times 15 \times 512$
$Conv2D + BN$		ReLU	$8 \times 15 \times 512$
$Maxpool$		–	$4 \times 15 \times 512$
FC		–	512
FC		ReLU	512
$Output$		Softmax	类别数

表 II 基于 RESNET 和 SERESNET 的非序列学生网络的架构详情。

层		激活函数	形状
Input		–	$257 \times 500 \times 1$
$Conv2D$		ReLU	$257 \times 500 \times 64$
$Maxpool$		–	$128 \times 250 \times 64$
$Conv2D + BN$ $Conv2D + BN$ $Conv2D + BN$	$\times 2$	ReLU ReLU ReLU	$128 \times 250 \times 96$
$Maxpool$		–	$64 \times 125 \times 96$
$Conv2D + BN$ $Conv2D + BN$ $Conv2D + BN$	$\times 3$	ReLU ReLU ReLU	$64 \times 125 \times 128$
$Maxpool$		–	$32 \times 62 \times 128$
$Conv2D + BN$ $Conv2D + BN$ $Conv2D + BN$	$\times 3$	ReLU ReLU ReLU	$32 \times 62 \times 256$
$Maxpool$		–	$16 \times 31 \times 256$
$Conv2D + BN$ $Conv2D + BN$ $Conv2D + BN$	$\times 3$	ReLU ReLU ReLU	$16 \times 31 \times 512$
$Maxpool$		–	$8 \times 15 \times 512$
FC		–	512
FC		ReLU	512
$Output$		Softmax	类别数

表 III 序列学生网络的架构详情。 N 表示输入被划分成的片段数量。

层	激活函数	形状
Input	–	$N \times 257 \times 100 \times 1$
$Encoder$	–	$N \times 512$
$BiLSTM$	ReLU	$N \times 512$
$BiLSTM$	ReLU	$N \times 512$
$Self - Attention$	注意力	512
$Output$	Softmax	类别数

如图所示3 （左）；(2) 在聚合器级进行视频信息蒸馏，如图 3 （右）所示。在编码器级蒸馏中，我们使用损失函数 $\mathcal{L}_E(E^S_{i_k}, E^T_{i_k})$ 来计算每个音频片段 $E^S_{i_k}$ 的嵌入与教师嵌入 $E^T_{i_k}$ 之间的损失，后者对应于该片段。随后，学生模型的编码器组件$F^{S_{enc}}$ 通过使用以下方式计算出的梯度进行训练

$$\nabla_s = (1-\alpha)\nabla\frac{\mathcal{L}_Y(y'_i, y_i)}{\nabla F^{S_{enc}}} + \alpha\sum_{k=0}^{M_i}\nabla\frac{\mathcal{L}_E(E^S_{i_k}, E^T_{i_k})}{\nabla F^S}, \quad (12)$$

而学生网络的聚合器和输出层仅使用由损失函数 $\mathcal{L}_Y(y'_i, y_i)$ 计算得到的梯度进行训练。

在聚合器级蒸馏中，我们使用损失函数 $\mathcal{L}_{AG}(E_i^{S_{agg}}, E_i^{T_{agg}})$，该函数计算学生模型的聚合器组件生成的嵌入与嵌入 $E_i^{T_{agg}}$ （使用公式5计算得到）之间的距离。学生网络的输出层使用以下方式进行训练：

仅 $\mathcal{L}_Y(y'_i, y_i)$，而其余流程则由使用以下方法计算出的梯度进行训练

$$\nabla_s = (1-\alpha)\nabla\frac{\mathcal{L}_Y(y'_i, y_i)}{\nabla F^{S_{agg}}} + \alpha\nabla\frac{\mathcal{L}_{AG}(E_i^{S_{agg}}, E_i^{T_{agg}})}{\nabla F^{S_{agg}}}. \quad (13)$$

C. 实现细节

我们解决方案的非序列版本中的网络由一个编码器和一个输出层组成。我们使用基于 VGG [38], ResNet [12], 和压缩激励（SE）网络 [16] 的 3 种架构，分别实现教师模型和学生网络。对于第一个基准测试，我们使用标准 VGG16 网络作为教师模型，并使用针对音频定制的基于 VGG 的网络作为学生网络。表 I列出了基于 VGG 的学生网络的详细信息。在该网络中，每个卷积层都与一个批归一化层相连。整个网络的卷积层和批归一化层均采用步长为 1。最大池化层采用 (2, 2) 的滤波器尺寸和步长 2。我们在最后一个最大池化层之后添加 2 个包含 512 个神经元的全连接层。最后一个全连接层的输出用作学生嵌入 E_s，如前文第 III节所述。卷积层和最后一个全连接层均使用线性整流单元（ReLU）作为激活函数。最后，任务头使用一个全连接层，其神经元数量等于实验中的类别数。然后对该层使用 Softmax函数。

对于第二个基准，我们在教师网络中采用标准的 ResNet34，并在学生网络中采用专为音频定制的基于 ResNet 的模型。基于 ResNet 的学生模型的详细信息见表II。学生模型的第一个卷积层使用尺寸为 (7, 7)、步长为 1 的滤波器。该层后接尺寸为 (2, 2)、步长为 2 的最大池化层。之后，我们在学生模型中使用残差块。每个残差块包含 3 组耦合的卷积层和批归一化层，以及一条将该块输入连接到输出的快捷连接。每个块中的第一个和最后一个卷积层使用尺寸为 (1, 1) 的滤波器，第二个卷积层使用尺寸为 (3, 3)、步长为 1 的滤波器。如表II所示，每个块会重复多次。我们在每个块后添加尺寸为 (2, 2)、步长为 2 的最大池化层。最后一个最大池化层后接 2 个全连接层，每层含有 512 个神经元；任务头中的全连接层所含神经元数等于类别数。与第一个基准类似，我们将 ReLU 用作卷积层和编码器中最后一个全连接层的激活函数，而任务头中的全连接层则使用 Softmax。

在第三个基准实验中，我们使用 SE 块构建网络。这些块采用与残差块相同的层格式，不同之处在于每个块都添加了一个 SE 模块。SE 模块由一个用于提取通道信息的全局池化层、两个中间带有 ReLU 激活函数的全连接层，以及位于全连接层之后的 sigmoid 激活函数组成。我们将标准 ResNet34 网络中的残差块替换为 SE 块，以构建教师网络。学生网络则通过将第二个基准实验中所用网络的残差块替换为 SE 块来构建。

在我们方案的序列版本中，网络由编码器组件和聚合器组件构成。与非序列格式类似，我们进行实验时使用了 3 个基准网络。对于教师模型和学生网络的编码器组件，我们分别采用非序列版本中使用的教师模型和学生网络。不过，学生网络的任务头被移除，编码器组件会为输入的每个片段生成一个大小为 512 的嵌入向量。网络的聚合器组件包含 2 个双向长短期记忆网络层，并配有一个注意力模块。表格III展示了所提方法序列版本实验中学生网络的详细信息。

IV. 实验

A. 数据集与数据准备

实验旨在评估仅进行音频表征学习与利用视频作为特权信息这两种方式在音频表征学习性能上的差异。实验针对语音识别和语音情感识别两项不同任务进行。我们使用了 3 个公开数据集，分别是**VoxCeleb** [31],瑞尔森情感语音与歌曲视听数据库（**RAVDESS**） [27], 和**IEMOCAP** [3]。

我们使用 VoxCeleb 数据集的音视频版本[31]进行语音识别。在这项任务中，我们旨在从一组已知说话人中识别给定话语的说话人。该版本的 VoxCeleb 数据集包含来自 1,211 人的 21,819 段音视频录音。我们将这 1,211 人所有录音中的 70% 用于训练，10% 用于验证，20% 用于测试。我们将音频的频谱图表征作为学生模型的输入。我们使用短时傅里叶变换（STFT）从音频中提取频率特征，窗口长度为 25 毫秒。该过程以 10 毫秒的窗口重叠在整个话语上重复进行。所有录音的话语时长并不相同。为避免输入可变长度带来的复杂性，训练模型时，我们将所有录音裁剪为 5 秒，从而得到尺寸为 257×500的频谱图。需要注意的是，推理时使用录音的原始时长。较短的话语通过重复填充，以匹配所需长度。视频以每秒 25 帧的帧率录制。我们使用自动人脸检测模型对每一帧进行标注，以确定说话者面部的位置和边界。数据集中边界的尺寸并不一致，因此我们裁剪图像并将其调整为固定尺寸 224×224。为评估所提方法的序列实现，我们将录音划分为 1 秒的片段。这样，每段话语会得到 5 个维度为 257×100 的较短频谱图序列，每段视频则会得到 5 组帧，每组包含 25 帧。

对于语音情感识别，我们使用两个数据集：RAVDESS [27] 和 IEMO-CAP [3]。在此任务中，我们旨在识别话语中说话人的情绪状态，并将其划分为不同的离散情绪类别，即悲伤、快乐、恐惧、厌恶、惊讶、愤怒和中性。RAVDESS 数据集由 24 名参与者的录音组成。我们采用留一被试交叉验证来训练和验证方法，并报告平均准确率。

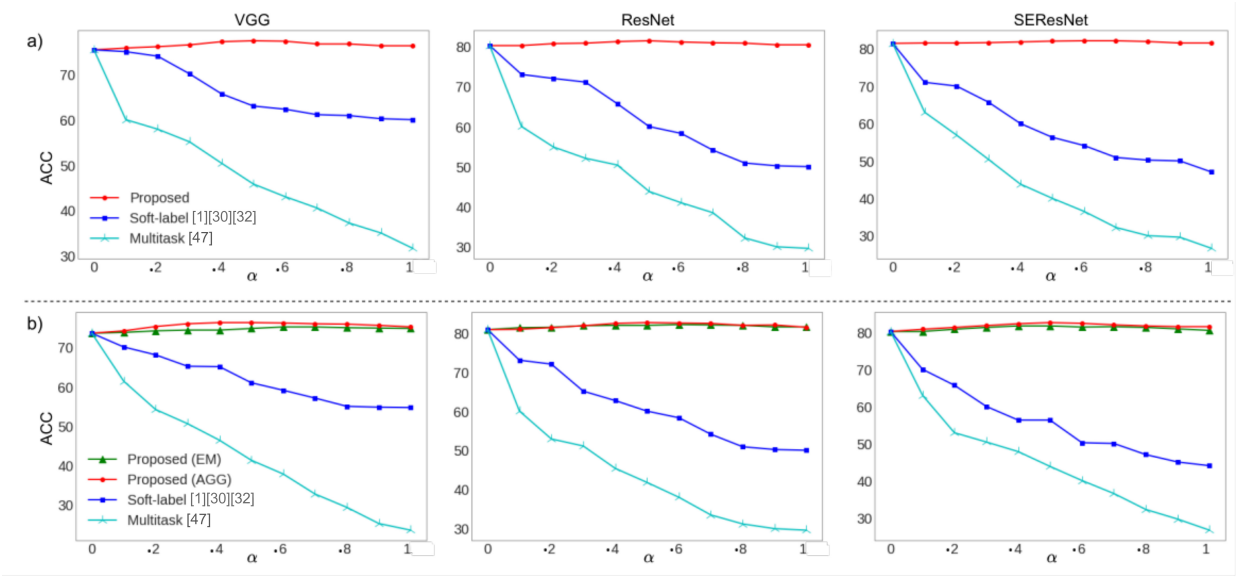

图 4. 比较在不同 α 值下，将所提方法用于特权信息学习与使用软标签和多任务训练进行说话人识别的效果：(a) 非序列设置；(b) 序列设置。

参与者需要以正常语音和歌唱两种形式，用 7 种不同情绪说出同一个句子。每位演员以正常语音表演该句子 60 次，以歌唱语音表演 44 次，其中一位演员不参与歌唱。我们的实验共使用 2,452 段视频录制。这些视频片段均在受控条件下录制，不含任何环境噪声。录音时长固定为 4 秒，因此无需对输入进行额外填充或裁剪。录制视频的帧率为每秒 25 帧。在大多数帧中，说话人的脸位于画面中央，且最多占画面的 50%。因此，我们对画面中央区域进行裁剪，并将所得图像调整为 224×224 的固定尺寸。

最后，我们还使用 IEMOCAP 评估所提方法在语音情感识别上的表现。该数据集包含由 10 人录制的共 6,000 段视听录制。我们采用 5 折交叉验证评估所提方法，并报告平均准确率。这些录音包含每位演员说出的单句即兴句子或脚本句子。每个话语由 3 位不同的人标注，并分为四种情绪类别：悲伤、快乐、愤怒和中性。数据集中录音的时长并不统一。因此，我们将录音时长固定为 4 秒：较长的话语通过截断处理，较短的话语则通过重复填充。

B. 训练细节

教师网络和学生网络在同一数据集上训练 50 个训练轮次。优化器使用 Adam 优化器[21]，$\beta 1 = 0.9$ 和 $\beta 2 = 0.99$。我们使用循环学习率 [39] 训练网络，初始学习率为 10^{-4}。我们选择循环学习率，以降低陷入局部极小值的概率。所有网络均在单个 Nvidia Titan RTX（24 GB 显存）图形处理器上以 32 的批量大小进行训练。

C. 基线方法

我们将所提方法与两种重新实现的基线进行比较：（1）“软标签蒸馏”，使用教师网络从视频模态生成的软标签来训练学生网络。这种方法已用于知识蒸馏研究，例如 [1], [30], [32]；在该基线中，参数 α （见公式 1）决定学生模型应在多大程度上遵循教师模型。（2）“多任务学习”，我们在第 II-C 节中对此进行了介绍。该方法此前已用于 [37]，以通过特权信息学习从视频中进行动作识别。在这种情况下，学生模型使用真实标签以及由次要解码器组件返回的梯度进行训练。该组件负责在次要模态的特征空间中生成训练样本的表示。此处参数α 表示赋予来自解码器组件的梯度的权重。

D. 性能

说话人识别。当特权信息融入框架后，我们使用识别准确率和等错误率来评估学生模型的性能。

我们排除了其他工作，例如 [26]，因为它们所提方法要求特权信息和原始训练数据来自相同模态。图4展示了使用所提方法后学生模型性能的实验结果，并将其与不同 α （模仿参数）取值和不同模型架构下的基线方法进行了比较。图中显示，对于所有 α取值，所提方法均能提升学生网络的性能。此外，我们观察到，随着 α 增大，基线方法会产生负面影响。这表明，尽管基线模型在使用视频提供

表 IV VoxCeleb 视频版本上的说话人识别实验结果。其中，(Acc_T)：教师网络的准确率；(Acc_S)：学生网络的准确率；(ΔAcc_S)：蒸馏后学生网络准确率的差值；EM：所提出的模型的嵌入级实现；AGG：所提出的模型的聚合级实现；所提出的模型的数值以 $\alpha = 0.5$ 报告。

模型	Acc_T	Acc_S $w/o\ Distill.$	Acc_S $w/\ Distill.$	ΔAcc_S%
所提出的模型（VGG）	78.4	75.5	77.5	2.64
所提出的模型（VGG+双向长短期记忆网络+EM）	80.4	73.7	75.3	2.17
所提出的方法（VGG+双向长短期记忆网络+AGG）	80.4	73.7	76.4	**3.66**
所提出的方法（ResNet）	83.1	80.3	81.5	1.49
所提出的方法（ResNet+双向长短期记忆网络+EM）	83.6	80.9	82.2	1.60
所提出的方法（ResNet+双向长短期记忆网络+AGG）	83.6	80.5	82.7	2.73
所提出的方法（SEResNet）	84.0	81.5	82.2	0.85
所提出的方法（SEResNet+双向长短期记忆网络+EM）	84.4	80.3	80.8	0.62
所提出的模型（SEResNet+双向长短期记忆网络+AGG）	84.4	80.5	82.7	2.73

表 V VoxCeleb1 标准测试集上的说话人验证实验结果。此处，(EER_S)：学生网络的等错误率；(ΔEER_S)：学生网络蒸馏后的等错误率变化值。EM：所提出模型的嵌入级实现；AGG：所提出模型的聚合级实现；所提出模型的数值以 $\alpha = 0.5$ 报告。

模型	EER_S $w/o\ Distill.$	EER_S $w/\ Distill.$	ΔEER_S%
所提出的模型（VGG）	10.15	9.7	4.43
所提出的模型（VGG+双向长短期记忆网络+EM）	9.34	8.12	13.06
提出的方法（VGG+双向长短期记忆网络+AGG）	9.34	8.06	13.70
提出的方法（ResNet）	5.33	3.91	**26.64**
提出的方法（ResNet+双向长短期记忆网络+EM）	5.36	4.05	24.44
提出的方法（ResNet+双向长短期记忆网络+AGG）	5.36	3.95	26.30
提出的方法（SEResNet）	5.31	4.22	20.52
提出的方法（SEResNet+双向长短期记忆网络+EM）	5.41	4.35	19.59
所提出的方法（SEResNet+双向长短期记忆网络+AGG）	5.41	4.27	21.07

真实标签缺失时为音频提供监督[1], [30], [32],方面表现出色，但在两种模态均参与监督训练、而推理时仅使用音频的场景中，并无任何益处。

还可以观察到，当 α 为 0.5 或 0.6 时，我们的方法影响最大。这表明，当特权信息蒸馏在网络训练中的权重与真实标签的权重几乎相等时，能够实现最佳性能提升。

表IV 展示了我们在 VoxCeleb 数据集的视频版本上，将所提方法应用于不同架构进行说话人识别的实验结果。我们给出了教师模型在 Acc_T 使用视频模态进行人物识别任务时的准确率。这使我们能够研究所提方法对不同性能教师模型所对应的学生模型的影响。我们还展示了未经特权信息蒸馏的学生网络的准确率，以便更好地观察采用所提方法带来的性能提升，并将其与其他方法进行比较。最后，我们展示了经过特权信息蒸馏后的学生网络准确率，并将其与未经蒸馏的学生模型进行比较（ΔAcc_S）。结果表明，采用所提方法后，学生网络的性能显著提升。在采用基于 VGG 的编码器的序列网络，并对聚合器进行特权信息蒸馏时，性能提升最高。此时，教师模型准确率与学生模型准确率之间的差值达到最大值，即 9.09%。我们还观察到，与教师网络和学生网络的性能差异相比，采用基于 ResNet 的编码器的序列网络并对聚合器进行特权信息蒸馏时，性能提升最大。

表 V 展示了我们在用于说话人验证的基准网络上进行的实验结果，测试数据来自 VoxCeleb1[31]标准测试集，其中包含来自训练集之外 40 位说话人的录音。我们以 EER_S 表示学生模型蒸馏前后的错误率，并通过计算差值并将其除以未进行蒸馏时的等错误率（$\Delta EER_S\%$），比较学生模型在这两个阶段的性能。我们观察到，采用非序列式实现时，基于 ResNet 的编码器错误率降幅最大，这表明与其他网络相比，这些编码器通常能从教师模型的训练中获益更多。

情绪识别。我们使用非加权准确率作为指标，评估采用我们的方法在 RAVDESS 上进行语音情感识别训练的网络的性能。对于以下所有取值 α，我们将我们的方法与前文介绍的重新实现的基线方法进行比较。图 5展示了这项实验的结果。我们发现，虽然我们的方法对学生网络产生了积极影响，但基线方法却产生了负面影响。这进一步表明，尽管利用视频

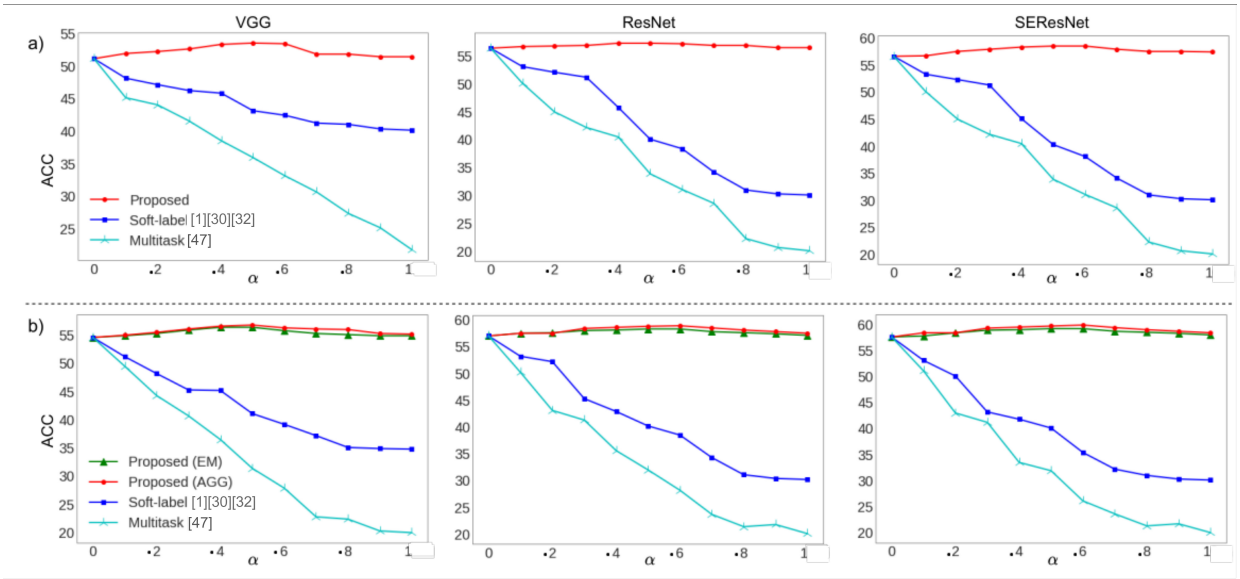

图 5. 在 RAVDESS 上进行语音情感识别时，对比所提方法用于特权信息学习与软标签和多任务训练的效果，以及不同 α 取值的影响：(a) 非序列设置；(b) 序列设置。

表 VI 语音情感识别（RAVDESS）的实验结果。此处，(Acc_T)：教师网络的准确率；(Acc_S)：学生网络的准确率；($Acc_T - Acc_S$)：蒸馏前教师网络与学生网络准确率之差；(ΔAcc_S)：蒸馏后学生网络准确率的变化；EM：所提出模型的嵌入级实现；AGG：所提出模型的聚合级实现；所提模型的数值以 $\alpha = 0.5$ 报告。

模型	Acc_T	Acc_S $w/o\ Distill.$	Acc_S $w\ Distill.$	ΔAcc_S%
所提方法（VGG）	80.5	51.0	52.3	2.54
所提方法（VGG+双向长短期记忆网络+EM）	83.4	54.6	56.4	3.29
所提方法（VGG+双向长短期记忆网络+AGG）	83.4	54.6	56.8	4.02
所提方法（ResNet）	81.1	56.4	57.3	1.59
所提方法（ResNet+双向长短期记忆网络+EM）	84.6	56.9	58.2	2.28
所提方法（ResNet+双向长短期记忆网络+AGG）	84.6	56.9	58.8	3.33
所提方法（SEResNet）	81.0	56.6	59.5	**5.12**
所提方法（SEResNet+双向长短期记忆网络+EM）	85.3	57.6	59.2	2.77
所提方法（SEResNet+双向长短期记忆网络+AGG）	85.3	57.6	59.9	3.45

作为学习音频的唯一监督来源、且没有真实标签时，基线模型是有效的；但在训练时使用两种模态、而推理时仅有音频可用的情况下，它们并不能提升深度神经网络的性能。

我们还通过比较所提方法集成到不同架构后的性能，进一步扩展了实验。表 VI 展示了这项评估的结果。我们列出了教师网络的准确率，以及蒸馏前后学生的准确率。可以看出，使用我们的方法后，学生模型的性能有所提升；采用基于SEResNet架构的非序列学生网络时，性能提升幅度最大。与基线方法比较时，我们也观察到了与先前实验相似的表现。

最后，我们在 IEMOCAP数据集上评估了所提方法。在这项实验中，我们旨在展示当教师模型的准确率低于学生网络时，所提方法的效果。我们将所提方法与前文所述的基线方法进行比较，涵盖α的所有取值。图6展示了这项实验的结果。如图所示，采用所提方法时，当α取值较低时，较弱教师模型不会对学生网络的性能产生负面影响；而在基线方法中，α取值很早就能观察到负面影响。表VII展示了将所提方法整合到不同架构后的性能。我们观察到，采用配备双向长短期记忆网络层的SEResNet架构时性能最佳；尽管教师模型的性能低于学生，所提方法并未对该网络的性能产生负面影响。

V. 结论与未来工作

我们采用教师-学生知识蒸馏进行特权信息学习（LUPI），以便在训练深度神经网络时同时利用音频和视频输入，而在推理时仅使用音频。在我们的框架中，首先使用教师模型从视频输入中提取嵌入。随后利用这些嵌入和真实标签来训练

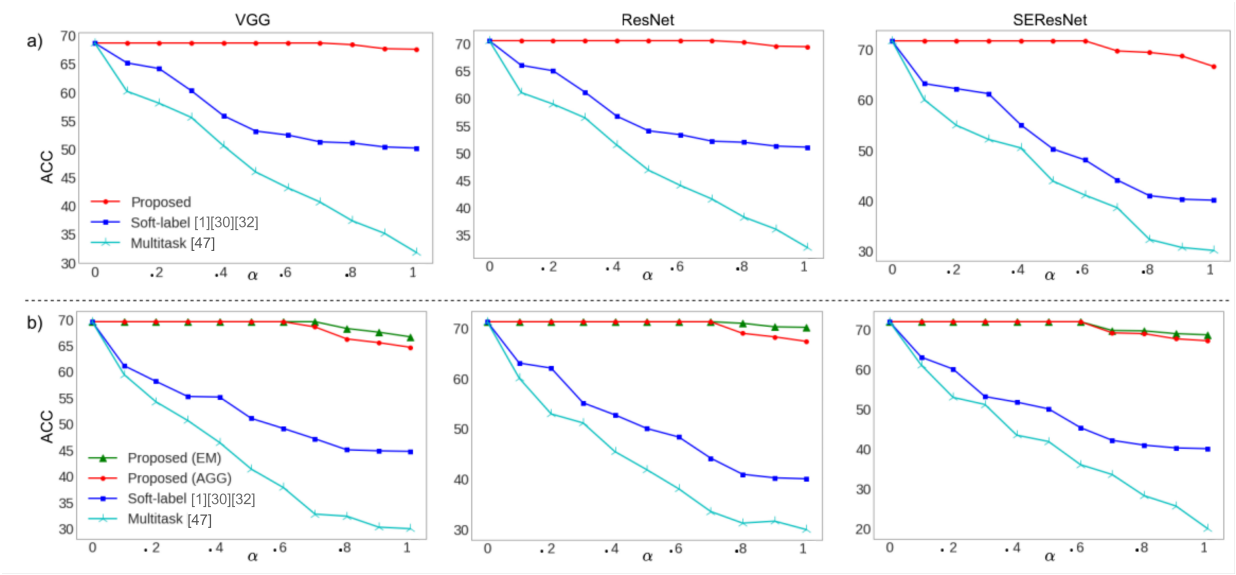

图 6. 在不同的 α 取值下，比较在 IEMOCAP 上将所提方法用于特权信息学习与使用软标签和多任务训练进行语音情感识别的效果：(a) 非序列设置；(b) 序列设置。

表 VII 语音情感识别实验结果（IEMOCAP）。其中，(Acc_T)：教师网络的准确率；(Acc_S)：学生网络的准确率；($Acc_T - Acc_S$)：蒸馏前教师网络与学生网络准确率之差；(ΔAcc_S)：蒸馏后学生网络准确率的变化；EM：所提出模型的嵌入级实现；AGG：所提出模型的聚合级实现；所提出模型的数值以 $\alpha = 0.5$ 报告。

模型	Acc_T	Acc_S $w/o\ Distill.$	Acc_S $w\ Distill.$	ΔAcc_S%
所提出的方法（VGG）	60.8	68.6	-25.65	0
所提出的方法（VGG+双向长短期记忆网络+EM）	64.3	69.6	-21.55	0
所提出的方法（VGG+双向长短期记忆网络+AGG）	64.3	69.6	-19.39	0
所提出的方法（ResNet）	62.3	70.2	-17.94	0
所提出的方法（ResNet+BiLSTM+EM）	65.2	70.9	-20.45	0
所提出的方法（ResNet+BiLSTM+AGG）	65.2	70.9	-18.33	0
所提出的方法（SEResNet）	62.0	71.3	-28.47	0
所提出的方法（SEResNet+BiLSTM+EM）	65.5	71.8	-23.95	0
所提出的方法（SEResNet+BiLSTM+AGG）	65.5	71.8	-21.86	0

学生网络。我们将该方法应用于非序列数据和序列数据两种不同设置。在非序列设置中，教师网络和学生网络均由编码器和任务头构成。我们使用教师模型编码器生成的嵌入来训练学生网络的编码器，而学生网络的任务头则使用真实标签进行训练。在序列设置中，我们在教师网络和学生网络中加入了一个额外的聚合组件，置于编码器与任务头之间。我们使用教师模型的编码器和聚合组件分别生成的两组嵌入，来训练学生网络的编码器和聚合组件。与非序列设置类似，学生网络的任务头使用真实标签进行训练。通过在语音识别（SR）和语音情感识别（SER）这两项任务上进行实验，我们表明，与以往研究相比，我们提出的框架能够显著提升学生网络的性能。尽管基准模型在语音识别和语音情感识别任务中依赖输入的不同方面，因此不同架构在各项任务上的性能也有所不同，但我们的方法始终提升了基准方法的性能。总而言之，我们的方法通过知识蒸馏，将特权信息学习融入利用视听数据进行的深度音频表示学习中，为其开辟了新的路径；在推理阶段仅使用音频即可。

我们的工作也为未来的研究提出了一系列新的挑战。未来研究可以立即着手的一步，是探索使用生成模型（例如生成对抗网络和变分自编码器）从教师模型中生成嵌入，以实现更好的领域适应和泛化。也可以在此背景下探索归一化流等较新的方法及即将出现的方法。

参考文献

[1] S. Albanie、A. Nagrani、A. Vedaldi 和 A. Zisserman。在真实场景中利用跨模态迁移进行语音情感识别。*ACM*国际多媒体会议，第 292–301 页，2018年。[2] B. T. Atmaja 和 M. Akagi。用于维度化视听情感识别的多任务学习与多阶段融合。*IEEE* 国际声学、语音与信号处理会议 *(ICASSP)*，第 4482–4486 页，2020年。

[3] C. Busso、M. Bulut、C.-C. Lee、A. Kazemzadeh、E. Mower、S. Kim、J. N. Chang、S. Lee 和 S. S. Narayanan。IEMOCAP：交互式情绪双人动作捕捉数据库。《语言资源与评估》，42(4)：335，2008年。[4] D. Chen、J.-P. Mei、Y. Zhang、C. Wang、Z. Wang、Y. Feng 和 C. Chen。带语义校准的跨层蒸馏。AAAI 人工智能大会，35(8)：7028–7036，2021年。[5] T. Do、T.-T. Do、H. Tran、E. Tjiputra 和 Q. D. Tran。用于视觉问答的紧凑三线性交互。IEEE/CVF 国际计算机视觉大会 (ICCV)，第 392–401 页，2019年。[6] M. Gao、Y. Wang 和 L. Wan。基于残差误差的知识蒸馏。《神经计算》，433：154–161，2021年。[7] N. C. Garcia、P. Morerio 和 V. Murino。用于动作识别的多流网络模态蒸馏。欧洲计算机视觉会议 (ECCV)，第 103–118 页，2018年。[8] Y. Guan、P. Zhao、B. Wang、Y. Zhang、C. Yao、K. Bian 和 J. Tang。用于知识蒸馏的可微特征聚合搜索。欧洲计算机视觉会议 (ECCV)，第 XXIV 部分，第 16 卷：469–484，2020年。[9] A. Hajavi 和 A. Etemad。用于短片段说话人识别的深度神经网络。INTERSPEECH，第 2878–2882 页，2019年。[10] A. Hajavi 和 A. Etemad。知道该听什么：用于深度语音表征学习的早期注意力。arXiv 预印本 arXiv:2009 01822，2020年。。[11] A. Hajavi 和 A. Etemad。用于真实场景端到端说话人识别的孪生胶囊网络。IEEE 国际声学、语音与信号处理会议 (ICASSP)，第 7203 7207 页，2021年 – 。[12] K. He、X. Zhang、S. Ren 和 J. Sun。《用于图像识别的深度残差学习》。IEEE/CVF 计算机视觉与模式识别大会 (CVPR)，第 770–778 页，2016年。[13] B. Heo、J. Kim、S. Yun、H. Park、N. Kwak 和 J. Y. Choi。特征蒸馏的全面革新。IEEE/CVF 国际计算机视觉大会 (ICCV)，第 1921–1930 页，2019年。[14] B. Heo、M. Lee、S. Yun 和 J. Y. Choi。通过蒸馏隐藏神经元形成的激活边界进行知识迁移。AAAI 人工智能大会，33(01)：3779–3787，2019年。[15] G. Hinton、O. Vinyals、J. Dean 等。蒸馏神经网络中的知识。arXiv 预印本 arXiv:1503.02531，2(7)，2015年。[16] J. Hu、L. Shen 和 G. Sun。挤压-激励网络。IEEE/CVF 计算机视觉与模式识别大会 (CVPR)，第 7132–7141 页，2018年。[17] Z. Huang 和 N. Wang。喜欢你所喜欢的：通过神经元选择性迁移进行知识蒸馏。arXiv 预印本 arXiv:1707.01219，2017年。[18] M. A. Jalal、E. Loweimi、R. K. Moore 和 T. Hain。利用胶囊路由学习用于语音情感识别的时间簇。INTERSPEECH，第 1701–1705 页，2019年。[19] I. Kansizoglou、L. Bampis 和 A. Gasteratos。在线视听情感识别的主动学习范式。《IEEE 情感计算汇刊》，1：1，2019年。[20] J. Kim、S. Park 和 N. Kwak。复杂网络改述：通过因子迁移实现网络压缩。神经信息处理系统国际会议 (NeurIPS)，第 2765–2774 页，2018年。[21] D. P. Kingma 和 J. Ba。Adam：一种随机优化方法。国际学习表征会议 (ICLR)，2015年。[22] N. Komodakis 和 S. Zagoruyko。更加关注注意力：通过注意力迁移提升卷积神经网络的性能。国际学习表征会议 (ICLR)，2017年。[23] T. Kourkounakis、A. Hajavi 和 A. Etemad。使用带双向长短期记忆网络的深度残差网络检测多种语音不流利现象。IEEE 国际声学、语音与信号处理会议 (ICASSP)，第 6089–6093 页，2020年。[24] T. Kourkounakis、A. Hajavi 和 A. Etemad。FluentNet：利用深度学习端到端检测口吃语音不流利现象。IEEE/ACM 音频、语音与语言处理汇刊，29：2986–2999，2021年。[25] P. Kumar、V. Kaushik 和 B. Raman。探索多模态语音情感识别的可解释性。INTERSPEECH，第 1748–1752 页，2021年。[26] J. Lambert、O. Sener 和 S. Savarese。利用异方差丢弃，在特权信息条件下进行深度学习。IEEE 计算机视觉与模式识别会议，第 8886–8895 页，2018年。[27] S. R. Livingstone 和 F. A. Russo。瑞尔森情感语音与歌曲视听数据库 (RAVDESS)：北美英语中面部和声音表达的动态多模态集合。《公共科学图书馆·综合》，13(5)：e0196391，2018年。[28] D. Lopez-Paz、L. Bottou、B. Sch¨olkopf 和 V. Vapnik。《统一蒸馏与特权信息》。国际学习表征会议（ICLR），第 1–10 页，2016年。[29] A. Nagrani、S. Albanie 和 A. Zisserman。《可学习的 PIN：用于人物身份的跨模态嵌入》。欧洲计算机视觉会议（ECCV），2018年。[30] A. Nagrani、J. S. Chung、S. Albanie 和 A. Zisserman。《利用跨模态自监督解耦语音嵌入》。IEEE 国际声学、语音与信号处理会议（ICASSP），第 6829–6833 页，2020年。[31] A. Nagrani、J. S. Chung、W. Xie 和 A. Zisserman。《VoxCeleb：真实环境下的大规模说话人验证》。《计算机语音与语言》，60:101027，2020年。[32] S. Nawaz、M. S. Saeed、P. Morerio、A. Mahmood、I. Gallo、M. H. Yousaf 和 A. Del Bue。《跨模态说话人验证与识别：多语言视角》。IEEE/CVF 计算机视觉与模式识别大会（CVPR）研讨会，第 1682–1691 页，2021年6月。[33] N. Passalis 和 A. Tefas。《利用概率知识迁移学习深层表征》。欧洲计算机视觉会议（ECCV），第 268–284 页，2018年。[34] P. Passban、Y. Wu、M. Rezagholizadeh 和 Q. Liu。《ALP-KD：基于注意力的层投影知识蒸馏》。AAAI 人工智能大会，35(15):13657–13665，2021年。[35] S. Roheda、B. S. Riggan、H. Krim 和 L. Dai。《跨模态蒸馏：条件生成对抗网络的应用》。IEEE 国际声学、语音与信号处理会议（ICASSP），第 2926–2930 页，2018年。[36] A. Romero、N. Ballas、S. E. Kahou、A. Chassang、C. Gatta 和 Y. Bengio。《FitNets：轻量深度网络的提示》。国际学习表征会议（ICLR），2015年。[37] Z. Shi 和 T.-K. Kim。《基于特权信息的 RNN 的学习与改进：从深度序列中进行动作识别》。IEEE/CVF 计算机视觉与模式识别大会（CVPR），第 3461–3470 页，2017年。[38] K. Simonyan 和 A. Zisserman。《用于大规模图像识别的超深卷积网络》。国际学习表征会议（ICLR），2015年。[39] L. N. Smith。《用于训练神经网络的循环学习率》。IEEE冬季计算机视觉应用会议（WACV），第 464–472 页，2017年。[40] F. M. Thoker 和 J. Gall。《用于动作识别的跨模态知识蒸馏》。IEEE图像处理国际会议（ICIP），第 6–10 页，2019年。[41] Y. Tian、D. Krishnan 和 P. Isola。《对比表征蒸馏》。arXiv 预印本 arXiv:1910.10699，2019年。[42] V. Vapnik 和 A. Vashist。《一种新的学习范式：利用特权信息学习》。《神经网络》，22(5-6):544–557，2009年。[43] W. Xie、A. Nagrani、J. S. Chung 和 A. Zisserman。《真实环境下说话人识别的话语级聚合》。IEEE 国际声学、语音与信号处理会议（ICASSP），第 5791–5795 页，2019年。[44] K. Xu、L. Rui、Y. Li 和 L. Gu。《用于图像分类的特征归一化知识蒸馏》。欧洲计算机视觉会议（ECCV），第 XXIV 部分，16:664–680，2020年。[45] S. You、C. Xu、C. Xu 和 D. Tao。《从多个教师网络中学习》。ACM知识发现与数据挖掘国际会议（SIGKDD），第 1285–1294 页，2017年。[46] C. Zhang 和 Y. Peng。《更好且更快：通过图蒸馏，从多个自监督学习任务中迁移知识以进行视频分类》。国际联合人工智能会议（IJCAI），第 1135–1141 页，2018年。[47] G. Zhou、Y. Fan、R. Cui、W. Bian、X. Zhu 和 K. Gai。《Rocket Launching：用于训练高性能轻量网络的通用高效框架》。AAAI 人工智能大会，2018年。