arXiv:2301.01456v1 [cs.CV] 4 Jan 2023

用于鲁棒语音识别的音视高效 Conformer

Maxime Burchi, Radu Timofte 计算机视觉实验室，CAIDAS，IFI，维尔茨堡大学¨，德国{maxime.burchi,radu.timofte}@uni-wuerzburg.de

摘要

近年来，基于神经网络的端到端自动语音识别（自动语音识别）系统取得了显著进步。大规模人工标注数据集的问世以及充足的计算资源，使训练强大的深度神经网络成为可能，并在学术基准测试中取得了极低的词错误率（词错误率）。然而，尽管模型在干净音频样本上表现出色，在噪声语音上的性能往往会下降。本文提出利用音频和视觉模态共同处理，提升近期提出的基于高效 *Conformer* 连接主义时序分类（*CTC*）的架构对噪声的鲁棒性。我们采用以 *ResNet-18* 为视觉前端、以 *Efficient Conformer* 为后端的结构，并在模块之间加入中间 *CTC* 损失，从而改进了以往的唇读方法。我们使用中间 *CTC* 残差模块，根据早期预测对中间块特征进行条件化，以放宽基于 *CTC* 的模型所采用的条件独立性假设。此外，我们用一种更高效、更简单的注意力机制取代 *Efficient Conformer* 的分组注意力，并将其称为分块注意力。我们在公开可用的唇读句子 *2*（*LRS2*）和唇读句子 *3*（*LRS3*）数据集上进行了实验。实验结果表明，使用音频和视觉模态有助于在环境噪声下更好地识别语音，并显著加快训练速度，仅用四分之一的训练步数即可取得更低的词错误率。我们的音频-视觉 *Efficient Conformer*（*AVEC*）模型达到了最先进的性能，在 *LRS2* 和 *LRS3* 测试集上的词错误率分别为 *2.3%* 和 *1.8%*。代码和预训练模型可在 *https://github.com/burchim/AVEC* 获取。

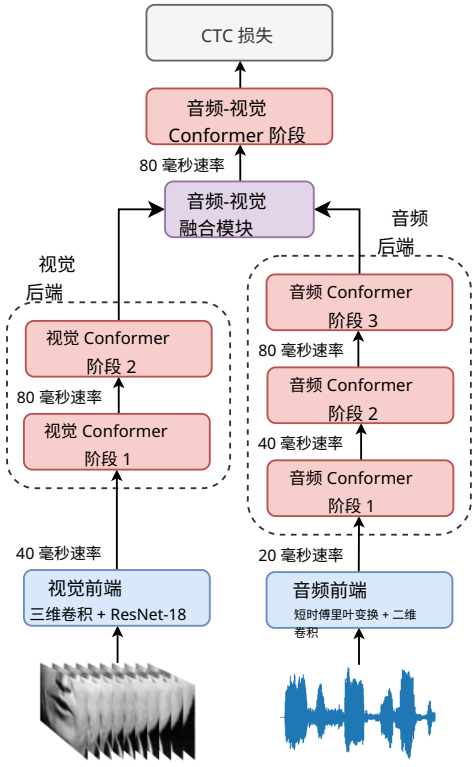


图 1：**音视高效 Conformer 架构**。该模型采用 CTC 损失进行端到端训练，并以说话人的原始音频波形和唇部动作为输入。

1. 引言

近年来，基于深度神经网络的端到端自动语音识别已成为最先进方法的标准 [25, 47, 18, 16, 17, 31, 7]。大规模人工标注数据集和充足计算资源的出现，使得训练强大的——

高效深度神经网络用于自动语音识别，在 LibriSpeech 等学术基准上取得了极低的词错误率 [34]。循环神经网络 (RNN) [15, 19]、卷积神经网络 (CNN) [10, 28] 和 Transformer [12, 23] 等神经网络架构，已成功利用原始音频波形和梅尔频谱图音频特征进行训练，将语音转录为文本。近期，Gulati 等人 [16]提出了一种卷积增强型 Transformer 架构（Conformer），通过卷积和注意力同时建模局部与全局依赖关系，从而提升语音识别性能。与此同时，Nozaki 等人 [33]

通过使用中间 CTC 损失，根据早期预测对中间编码器块特征进行条件化，从而改进了基于CTC的语音识别 [14]。Burchi 等人 [7]还提出了一种用于语音识别的高效 Conformer 架构，采用分组注意力，在降低计算量的同时提升了性能。受计算机视觉骨干网络的启发，高效 Conformer 编码器由多个阶段组成，每个阶段包含若干 Conformer 块，逐步对音频序列进行下采样，并将其投影到更宽的特征维度。

然而，即使这些仅音频方法已经打破了最先进水平，它们在现实世界中应用时的一大缺陷仍是在存在环境噪声时性能会迅速下降。与此同时，视听语音识别（AVSR）近期因能够利用图像处理技术辅助语音识别系统而受到广泛关注。先前的研究表明，纳入唇部动作这一视觉模态，可以提高自动语音识别系统对噪声的鲁棒性，并取得更好的识别性能 [41, 42, 36, 1, 45, 29]。Xu 等人 [45] 提出了一种两阶段方法：首先利用说话人的唇部动作将目标语音与背景噪声分离，然后借助唇部动作转写滤波后的音频信号。Petridis 等人 [36] 采用混合架构，通过早期融合策略训练带有辅助CTC损失的基于LSTM的序列到序列（S2S）模型，以获得更好的性能。Ma 等人 [29]采用配备ResNet-18 [20] 前端网络的Conformer后端网络，以提升识别性能。

其他研究聚焦于视觉语音识别（VSR），仅利用唇部动作将口语转写为文本 [4, 9, 48, 3, 49, 37, 30]。一项重要的研究方向是采用跨模态蒸馏。Afouras *etal.* [3] 和 Zhao *et al.* [49] 提出，通过从在大规模纯音频语料库上训练的自动语音识别模型中进行蒸馏，提升唇读性能；而 Ma*et al.* [30]则采用基于预测的辅助任务。Prajwal*et al.* [37]还提出使用子词单元而非字符来转写序列，大幅降低运行时间和内存需求，同时提供语言先验，减轻模型的语言建模负担。

本文聚焦于设计一种同时处理音频和视觉模态的抗噪语音识别架构。我们采用近期提出的基于 CTC 的 Efficient Conformer 架构 [7]，并证明，加入唇部动作这一视觉模态，能够有效提升噪声鲁棒性，同时显著加快训练速度。与仅音频版本相比，我们的音视高效 Conformer（AVEC）以少 4 倍的训练步数取得了更低的词错误率。此外，我们是首个在模块之间应用中间 CTC 损失的工作，以提升视觉语音识别性能。我们证明，利用中间 CTC 残差模块，根据早期预测对中间特征进行条件化，可以缩小基于序列到序列（S2S）的自回归与非自回归音视频语音识别系统之间的词错误率差距。这也有助于应对一种常见的失败情形：视听模型往往会忽略视觉模态。通过这种方式，我们促使融合前层学习时空特征。最后，我们用一种更高效、更简单的注意力机制取代 Efficient Conformer 的分组注意力，并将其称为块注意力。块注意力的性能与分组注意力相近，但复杂度更低。本文的贡献如下：

- 我们通过同时处理音频和视觉模态，提升了最新提出的 Efficient Conformer 架构的噪声鲁棒性。
- 我们使用中间 CTC 残差模块，根据早期预测对中间 Conformer 块特征进行条件化，从而放宽 CTC 模型的条件独立假设。这使我们能够缩小基于序列到序列（S2S）的自回归方法与非自回归方法之间的词错误率差距。
- 我们提出用一种更高效、更简单的注意力机制来替代 Efficient Conformer 的分组注意力，并将其称为块注意力。与分组注意力相比，块注意力在复杂度更低的同时，性能相近。
- 我们在公开的 LRS2 和 LRS3 数据集上进行了实验，并使用音频和视觉模态取得了最先进的结果。

2. 方法

本节将介绍我们提出的音视高效 Conformer 网络。该模型由 4 个主要组件构成：音频编码器、视觉编码器、音视融合模块和视听编码器。音频和视觉编码器分别由模态专用前端网络和 Efficient Conformer 后端网络构成，前者将各输入模态转换为时间序列，后者则用于建模局部和全局时间关系。整个模型采用端到端方式训练，除输出 CTC 层外，还在 Conformer 块之间使用中间 CTC 损失。模型的完整架构如图 1 所示。

2.1. 模型架构

音频前端。音频前端网络首先使用短时傅里叶变换将原始音频波形转换为梅尔频谱图，变换以 20 ms 的窗口进行，步长为 10ms。随后得到 80 维

频率特征，并应用梅尔尺度对数滤波器组。梅尔频谱图经二维卷积干网络处理，以提取局部时频特征，生成帧率为 20ms 的信号。音频前端的架构如表 1 所示。

表 1：音频前端架构，共 120 万个参数。 T_a 表示输入音频样本长度。

阶段	层	输出形状
傅里叶变换	短时傅里叶变换：窗长 400 帧移长度 160，512 个快速傅里叶变换	(257， $T_a//160+1$)
Mel 缩放	梅尔尺度：80 个梅尔频带	(80， $T_a//160+1$)
Stem	二维卷积： 3^2, 180 个滤波器， 2^2 步幅	(180, 40, $T_a//320+1$)
Proj	线性层，180 个单元	($T_a//320+1$，180)

视觉前端。视觉前端网络 [29]将输入视频帧转换为时间序列。首先对视频应用卷积核大小为 $5\times7\times7$ 的三维卷积主干。随后，使用带有空间平均池化输出的二维 ResNet-18 [20]，对每个视频帧进行独立处理。接着，使用线性层将时间特征投影到后端网络的输入维度。视觉前端架构如表 2 所示。

表 2：视觉前端架构，共 1130 万个参数。 T_v 表示输入视频帧的数量。

阶段	层	输出形状
Stem	三维卷积： 5×7^2， 64 个滤波器， 1×2^2 步幅 三维最大池化： 1×3^2, 1×2^2 步幅	(64， T_v, 22， 22)
残差块 1	$2\times$ [二维卷积： 3^2, 64滤波器 二维卷积： 3^2, 64滤波器]	(T_v,64,22， 22)
残差块 2	$2\times$ [二维卷积： 3^2, 128个滤波器 二维卷积： 3^2, 128个滤波器]	(T_v， 128， 11， 11)
残差块 3	$2\times$ [二维卷积： 3^2, 256个滤波器 二维卷积： 3^2, 256个滤波器]	(T_v， 256， 6， 6)
残差块 4	$2\times$ [二维卷积： 3^2, 512个滤波器 二维卷积： 3^2, 512个滤波器]	(T_v， 512， 3， 3)
Pool	全局平均池化	(T_v， 512)
Proj	线性层，256个单元	(T_v， 256)

后端网络。后端网络采用 Efficient Conformer 架构。高效 Conformer 编码器由若干阶段组成，每个阶段包含多个 Conformer 块，并采用带有相对位置编码的分组注意力。时间序列通过步幅卷积逐步下采样，并投影到更宽的特征维度，从而降低计算量，同时取得更好的性能。我们在音频后端网络中使用 3 个阶段，将音频信号下采样至 80 毫秒的帧率。只需 2 个阶段，即可将视觉信号下采样至相同的帧率。表 6 展示了各后端网络的超参数。

表 3：后端网络超参数。*InterCTC* 块表示带有后置 Inter CTC 残差模块的 Conformer 块。

网络	视觉后端	音频后端	音频-视觉编码器
参数数量（M）	13.6	17.9	15.9
阶段数	2	3	1
每阶段模块数	6, 1	5, 6, 1	5
总块数	7	12	5
阶段特征维度	256, 360	180, 256, 360	360
卷积核大小	15	15	15
阶段补丁大小	1, 1	3, 1, 1	1
InterCTC 块	3, 6	8, 11	2

音视融合模块。与 [36, 29], 类似，我们采用早期融合策略来学习视听特征并降低模型复杂度。后端网络输出的声学和视觉特征会被拼接，并输入联合前馈网络。大小为 $2\times d_{model}$ 的拼接特征首先通过输出大小为 $d_{ff}=4\times d_{model}$ 的线性层进行扩展，再经过 Swish 激活函数 [38]，然后投影回原始特征维度 d_{model}。

音频-视觉编码器。音频-视觉编码器是一个单阶段后端网络，由 5 个未进行下采样的 Conformer 块组成。随后，编码器的输出被投影到 CTC 层，以最大化正确目标对齐的概率之和。

2.2. 分块注意力。

Efficient Conformer [7] 提出在较早的编码器层中用分组注意力替代多头自注意力（MHSA） [44] 。分组 MHSA 会先沿特征维度对相邻的时间元素进行分组，再应用缩放点积注意力，从而降低注意力复杂度。由于注意力的计算复杂度与序列长度的平方成正比，这会导致网络的复杂度不对称：较早的注意力层所需的浮点运算次数多于序列长度较短的后续层。在本研究中，我们提出用一种更简单、更高效的注意力机制替代分组注意力，并将其称为块注意力（图 2）。与多尺度视觉变换器（MViT） [13]为视频和图像识别提出的池化注意力类似，块注意力会先对输入序列进行平均

表4：注意力变体的复杂度，包括查询、键、值和输出线性投影。 n 和 d 分别表示序列长度和特征维度。

注意力变体	超参数	全注意力复杂度
常规	-	$O(n\cdot d^2+n^2\cdot d)$
分组	组大小（g）	$O(n\cdot d^2+(n/g)^2\cdot d\cdot g)$
分块	块大小（k）	$O(n/k\cdot d^2+(n/k)^2\cdot d)$

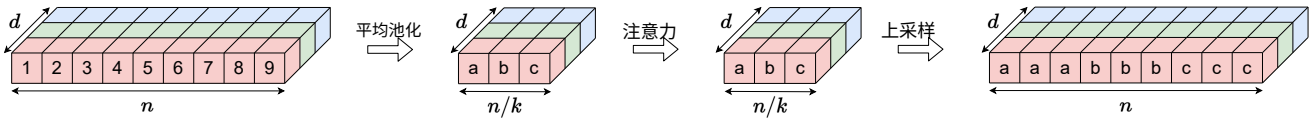


图 2：**分块多头自注意力**。 输入序列在应用多头自注意力之前，先通过平均池化进行下采样。随后通过最近邻上采样对输出序列进行上采样，将注意力复杂度从 $O(n^2 \cdot d)$ 降低至 $O((n/k)^2 \cdot d)$，其中 k 定义了池化/上采样核大小。当 $k = 1$时，块注意力等同于常规注意力。

池化，然后再投影查询、键和值。

$$X = AvgPooling1d(X_{in}) \tag{1}$$

$$\text{with } Q, K, V = XW^Q, XW^K, XW^V \tag{2}$$

其中， W^Q、 W^K、 $W^V \in \mathbb{R}^{d\times d}$分别为查询、键和值的线性投影参数矩阵。随后，在较低分辨率下执行带有相对正弦位置编码的多头自注意力：

$$MHSA(X) = Concat\left(O_1, ..., O_H\right) W^O \tag{3}$$

$$\text{with } O_h = softmax\left(\frac{Q_h K_h^T + S_h^{rel}}{\sqrt{d_h}}\right) V_h \tag{4}$$

其中， $S^{rel} \in \mathbb{R}^{n\times n}$ 是相对位置得分矩阵，满足 $S^{rel}[i,j] = Q_i E_{j-i}^T$。 E 是标准正弦位置编码矩阵的线性投影，其位置范围从 $-(n_{max}-1)$到（$n_{max}-1$）。随后，对注意力输出序列进行投影，并通过最近邻上采样将其恢复到初始分辨率。

$$X_{out} = UpsampleNearest1d(MHSA(X)) \tag{5}$$

因此，同一分块中的每个时间元素都会产生相同的注意力输出。局部时间关系仅由卷积模块建模，而全局关系则由块注意力建模。本研究采用一维块，但块注意力

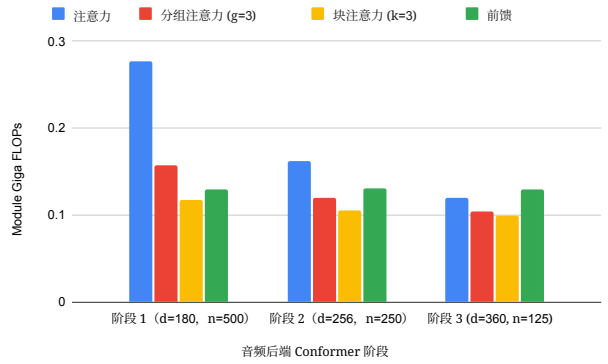


图3： **仅音频后端模块的浮点运算次数（十亿）**。

也可以通过使用二维和三维块推广到图像和视频数据。我们将此留待未来研究。表4展示了各注意力变体的计算复杂度。与分组注意力相比，路径注意力通过减少查询、键、值和输出全连接层所需的计算量，进一步降低了复杂度，同时保持特征维度不变。与以往工作类似 [7], ，我们仅在第一个音频后端阶段使用块注意力，以降低复杂度，同时保持模型的识别性能。图3展示了各注意力模块变体的浮点运算次数随编码序列长度 n和模型特征维度 d变化的情况。在第一个音频后端阶段，使用块注意力或分组注意力变体取代常规多头自注意力，可大幅减少浮点运算次数。

2.3. 中间CTC预测。

受 [27] 和 [33]启发，他们提出在编码器块之间添加中间 CTC 损失，以提升基于 CTC 的语音识别性能。我们在编码器网络中加入中间 CTC 残差模块（图4）。我们以早期预测为条件，对音频、视觉和音频-视觉编码器的中间块特征进行处理，从而放宽 CTC 模型的条件独立性假设。在训练和推理过程中，每个中间预测都会与下一层的输入相加，以辅助识别。我们采用 [33] 提出的相同方法，但不在各项损失之间共享层参数。 l^{th} 块的输出 X_l^{out}经过带有残差连接和 Softmax激活函数的前馈网络处理：

$$Z_l = Softmax(Linear(X_l^{out})) \tag{6}$$

$$X_{l+1}^{in} = X_l^{out} + Linear(Z_l) \tag{7}$$

其中， $Z_l \in \mathbb{R}^{T\times V}$ 是输出词汇表上的概率分布。随后，使用目标序列 y 计算中间 CTC 损失，计算方式如下：

$$L_l^{inter} = -log(P(y|Z_l)) \tag{8}$$

$$\text{with } P(y|Z_l) = \sum_{\pi\in\mathcal{B}_{CTC}^{-1}(y)} \prod_{t=1}^{T} Z_{t,\pi_t} \tag{9}$$

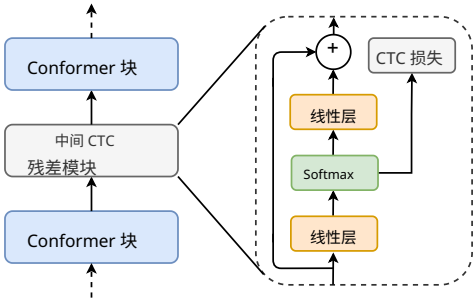


图 4：**中间 CTC 残差模块**。中间预测结果会与下一个 Conformer 块的输入相加，使最终模块的预测受到该结果的调节。中间 CTC 损失会与输出 CTC 损失相加，用于计算最终损失。

其中，$\pi \in V^T$ 是词元路径，$\mathcal{B}_{CTC}$ 是一个多对一映射，它只需从路径中移除所有空白符和重复标签。总训练目标定义如下：

$$L = (1-\lambda)L^{CTC} + \lambda L^{inter} \tag{10}$$

$$\text{with } L^{inter} = \frac{1}{K}\sum_{k \in interblocks} L_k^{inter} \tag{11}$$

其中，*interblocks*是包含后置 Inter CTC 残差模块的模块集合（图 4）。与 [33] 类似，我们在每 3 个 Conformer 块后使用中间 CTC 残差模块，并在所有实验中将 λ 设为 0.5。

3. 实验

3.1. 数据集

本研究使用了 3 个公开可用的音视频语音识别数据集。野外唇读（LRW） [8] 数据集用于视觉预训练，而唇读句子 2（LRS2） [1] 和唇读句子 3（LRS3） [2] 数据集用于训练和评估。

LRW 数据集。LRW 是一个音频-视觉单词识别数据集，包含一系列短视频片段，每个片段中只有一个单词，词汇表共包含 500 个词。该数据集包含 488,766 个训练样本，每个类别至少有 800 个话语；验证集和测试集各包含 25,000 个样本，每个类别有 50 个话语。

LRS2 和 LRS3 数据集。LRS2数据集包含来自英国广播公司电视节目的 224.1 小时视频，共有 144,482 个片段；LRS3数据集则包含从 TED 和 TEDx 演讲中提取的 438.9 小时视频，共有 151,819 个片段。两个数据集均包含对应的字幕及词语对齐边界，并由预训练划分、训练验证划分和测试划分组成。LRS2 的预训练部分有 96,318 段话语（195 小时），训练部分有 45,839 段（28 小时），验证部分有 1,082 段（0.6 小时），测试部分有 1,243 段（0.5 小时）。LRS3 的预训练集有 118,516 段话语（408 小时），训练验证集有 31,982 段话语（30 小时），测试集有 1,321 段话语（0.9 小时）。所有视频均只包含一位说话人，分辨率为 224×224 像素，采样帧率为 25 帧每秒，音频采样率为 16 kHz。

3.2. 实现细节

预处理 与 [29], 类似，我们通过边界框裁剪唇部区域，以消除旋转和缩放带来的差异；边界框尺寸为 96×96 像素，以便于识别。我们使用 RetinaFace [11] 人脸检测器和人脸对齐网络（FAN） [6]检测 68 个面部关键点。随后将裁剪后的图像转换为灰度图，并归一化至 -1 到 1 之间。LRW、LRS2 和 LRS3 数据集的面部关键点取自先前的工作 [30]，并用于预处理，以便对各方法进行公平比较。我们使用 SentencePiece [26]，根据 LRS2&3 预训练集以及训练集和验证集划分中的转录文本构建字节对编码分词器。遵循以往基于 CTC 的语音识别研究 [31, 7]，我们使用大小为 256 的词汇表，其中包括 CTC 空白标记。

数据增强训练期间，我们对音频梅尔频谱图应用 SpecAugment [35]，以防止过拟合：使用两个频率掩码，掩码大小参数为 $F = 27$，以及五个大小自适应的时间掩码 $pS = 0.05$。与 [30], 类似，我们沿时间轴对视频进行掩码，每秒使用一个掩码，最长掩码时长设为 0.4 秒。训练期间还会对每段视频进行大小为 88×88 的随机裁剪和水平翻转。对于纯视觉实验，我们还遵循 Prajwal 等人 [37]的方法，在测试时采用中心裁剪和水平翻转。

训练设置我们首先在 LRW 数据集上使用交叉熵损失对视觉编码器进行预训练，使其能够识别说出的词语。视觉编码器经过 30 个训练轮次的预训练，随后将前端权重用作训练的初始化参数。我们使用 Noam 调度策略，在 LRS2&3 数据集上训练音频和视觉编码器，预热步数为 10k，峰值学习率为 1e-3。我们使用 Adam 优化器 $\beta_1 = 0.9, \beta_2 = 0.98$。此外，还对模型的所有可训练权重添加权重为 1e-6 的 L2 正则化。我们在 4 个 GPU 上以 256 的全局批次大小训练所有模型，每个 GPU 的批次大小为 16，并累积 4 步。仅视觉和音频-视觉实验使用 Nvidia A100 40GB GPU，音频单模态实验则使用 RTX 2080 Ti。纯音频模型训练 200 个轮次，而仅视觉模型和视听模型分别训练 100 和 70 个轮次。请注意，训练期间我们只保留长度少于 400 帧（16 秒）的视频。最后，在评估之前，我们使用随机权重平均对最后 10 个训练轮次检查点的模型权重进行平均。

表5：采用公开和非公开数据集的最新方法在 LRS2/LRS3 测试集上的词错误率（%）比较，涉及仅音频（音频输入）、仅视觉（视觉输入）和音频-视觉（音视频）模型。

方法	模型准则	训练数据集	总计小时	测试词错误率 AO	VO	AV
（↓） 使用公开可用的数据集（↓）						
Petridis 等人 [36]	CTC+序列到序列	LRW、LRS2	381	8.3 / -	63.5 / -	7.0 / -
张等 [48]	S2S	LRW、LRS2&3	788 / 790	-	51.7 / 60.1	-
Afouras 等人 [3]	CTC	VoxCeleb2^{clean}，LRS2&3	1,032 / 808	-	51.3 / 59.8	-
Xu 等人 [45]	S2S	LRW、LRS3	595	- / 7.2	- / 57.8	- / 6.8
Yu等人[46]	LF-MMI	LRS2	224	6.7 / -	48.9 / -	5.9 / -
Ma等人 [29]	CTC+序列到序列	LRW、LRS2&3	381 / 595	3.9 / 2.3	37.9 / 43.3	3.7 / 2.3
Prajwal等 [37]	S2S	LRS2&3	698	-	28.9 / 40.6	-
Ma等 [30]	CTC+序列到序列	LRW、LRS2&3	818	-	**27.3 / 34.7**	-
Ours	CTC	LRW, LRS2&3	818	**2.8 / 2.1**	32.6 / 39.2	**2.5 / 1.9**
+ 神经语言模型	CTC	LRW, LRS2&3	818	**2.4 / 2.0**	29.8 / 37.5	**2.3 / 1.8**
(↓) 使用非公开数据集(↓)						
Afouras 等人 [1]	S2S	MVLRS、LRS2&3	1,395	9.7 / 8.3	48.3 / 58.9	8.5 / 7.2
赵等人 [49]	S2S	MVLRS，LRS2	954	-	65.3 / -	-
Shillingford 等人 [40]	CTC	LRVSR	3,886	-	- / 55.1	-
Makino等人 [32]	转导器	YouTube-31k	31,000	- / 4.8	- / 33.6	- / 4.5
Serdyuk 等人 [39]	转导器	YouTube-90k	91,000	-	- / 25.9	- / 2.3
Prajwal等 [37]	S2S	MVLRS、TEDx$_{ext}$、LRS2&3	2,676	-	22.6 / 30.7	-
Ma等 [30]	CTC+序列到序列	LRW、AVSpeech、LRS2&3	1,459	-	25.5 / 31.5	-

语言模型。与 [28], 类似，我们使用 N 元语法统计语言模型（LM）和 Transformer 神经语言模型进行实验。我们使用 6 元语法语言模型通过集束搜索生成一组假设，并使用外部 Transformer 语言模型对最终列表进行重评分。6 元语法语言模型在 LRS2&3 预训练集和训练验证转录文本上训练。对于神经语言模型，我们在 LibriSpeech 语言模型语料库上对 12 层 GPT-3 Small [5]进行 0.5M 步预训练，批次大小为 0.1M 词元，然后在 LRS2&3 转录文本上对其微调 10 个训练轮次。

3.3. 结果

表5比较了我们的音频-视觉高效 Conformer 与当前最先进方法在 LRS2 和 LRS3 测试集上的词错误率。我们的音频-视觉高效 Conformer 取得了当前最先进的表现，词错误率为 2.3%/1.8%。在仅视觉赛道中，我们的 CTC 模型与采用序列到序列准则的最新自回归方法不相上下。我们能够复现相近的结果，但仍稍逊于*Ma* 等人 [30]，后者使用了辅助损失以及预训练的纯音频和仅视觉网络。我们发现，与音频单模态实验相比，我们的音视网络收敛更快，仅用四分之一的训练步数便达到了更好的性能。视觉编码器的中间 CTC 损失能够降至低于仅视觉实验的水平，这表明优化音视层有助于融合前层学习更好的表征。

3.4. 消融研究

我们提出了详细的消融研究，以更好地了解各项架构改动在复杂度和词错误率方面带来的提升。我们报告了网络处理一个十秒音频/视频片段所需的操作次数，以浮点运算次数（乘法和加法的次数）衡量。我们还在 LRS3 测试集上测量了逆实时因子，解码时批次大小设为 1，并使用单个 Intel Core i7-12700 CPU 线程。所有消融实验均通过训练纯音频模型 200 个训练轮次，以及仅视觉/视听模型 50 个训练轮次完成。

高效 Conformer 视觉后端。我们通过采用高效 Conformer 后端网络，改进了近期提出的视觉 Conformer 编码器 [29] 。使用字节对编码进行分词，而非使用字符，使我们能够进一步对时间序列进行下采样，同时不影响 CTC 损失的计算。表 6 显示，在我们的仅视觉模型中采用高效 Conformer 后端网络，能够提升性能，同时降低模型复杂度和训练时间。模型参数数量也略有减少。

表 6：视觉后端网络的消融研究。

视觉后端	参数量（百万）	LRS2 test	LRS3 test	#浮点运算次数（十亿）	Inv RTF
Conformer	43.0	39.53	47.14	87.94	5.17
Eff Conf	40.4	**37.39**	**44.96**	84.52	5.26

参考

作者们查看了 10 年间发表的论文，其中有数百篇不得不被剔除

输出

区块 3： 这 些作者查阅了论文，历时数年，数百篇论文最终被丢弃和 各国的情况也大致如此区块 6：这些 作者查阅了论文，历时数年 时间跨度 并 数百篇 论文都被丢弃区块 9：这些 作者查阅了论文 撰写于 历时 八 年 期间 数百篇篇 论文 都被丢弃 了区块 12：作者查阅了论文 撰写于 历时 10 年，期间有数百篇论文不得不 被 丢弃 了

图 5：在 LRS3 测试集上使用贪心搜索解码，并每 3 个模块进行一次中间 CTC 预测时，我们的仅视觉模型的输出示例。除了在“10 year”前漏掉了“a”之外，该句几乎被正确转录。

中间 CTC 残差模块。与 [33], 类似，我们尝试在模块之间添加中间 CTC 残差模块，以放宽 CTC 的条件独立假设。表 7 显示，每 3 个 Conformer 块使用中间 CTC 损失，能显著降低词错误率；但在纯音频设置下并未提升性能。图 5 展示了一个示例：在 LRS3 测试集上，不使用外部语言模型，通过贪心搜索对中间模块的预测进行解码。可以看到，编码器层通过以前一层的中间预测为条件，逐步优化输出。由于我们的模型基于帧级预测逐步优化输出，因此除了替换错误之外，还能纠正插入错误和删除错误。我们还通过测量音频-视觉模型在两种模态之一被屏蔽时的性能，进一步研究中间 CTC 对多模态学习的影响。正如先前的研究所指出的， [8, 1, 32], 具有多模态输入的网络往往会被其中一种模式主导。在我们的案例中，语音识别远比唇读容易，这可能导致模型忽略视觉信息。表 8 显示，中间 CTC 可以通过迫使融合前层转录输入信号来帮助应对这一问题。

表 7：中间 CTC 残差模块的消融研究。

模型 后端	#参数量 （百万）	LRS2 test	LRS3 test	#浮点运算次数 （十亿）	Inv RTF
仅音频 (↓)					
Eff Conf	31.5	2.83	2.13	7.54	51.98
+ 中间 CTC	32.1	2.84	2.11	7.67	50.30
仅视觉 (↓)					
Eff Conf	40.4	37.39	44.96	84.52	5.26
+ 中间 CTC	40.9	**33.82**	**40.63**	84.60	5.26
音频-视觉 (↓)					
Eff Conf	60.9	2.87	2.54	90.53	4.84
+ 中间 CTC	61.7	2.58	1.99	90.66	4.82

表8：在模态掩蔽设置下，中间 CTC 对 LRS2 / LRS3测试集上视听模型词错误率（%）的影响。

中间 CTC	视听评估模式 掩蔽视频	掩蔽音频	无掩蔽
No	4.48 / 3.22	52.77 / 59.10	2.87 / 2.54
Yes	**3.39 / 2.38**	**37.62 / 46.55**	**2.58 / 1.99**

分块多头自注意力。我们尝试在第一个音频编码器阶段用块注意力替代分组注意力，旨在提高模型效率并简化模型，同时不损害性能。分组注意力 [7] 旨在降低第一个编码器阶段中长序列的注意力复杂度。表 9 展示了各类注意力变体对仅音频模型性能和复杂度的影响。我们从 Efficient Conformer (M) [7]开始，并替换其注意力机制。我们发现，在第一个后端阶段使用块大小为 3 的分块注意力，可以替代分组注意力而不损失性能。

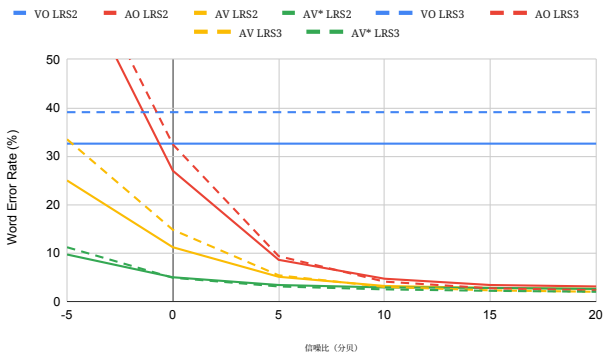
表 9：音频后端注意力的消融研究。

注意力 Type	组 / 块大小	LRS2 test	LRS3 test	#浮点运算次数 （十亿）	Inv RTF
常规	-	2.85	2.12	8.66	49.86
分组	3, 1, 1	2.82	2.13	8.06	50.27
分块	3, 1, 1	2.83	2.13	**7.54**	**51.98**

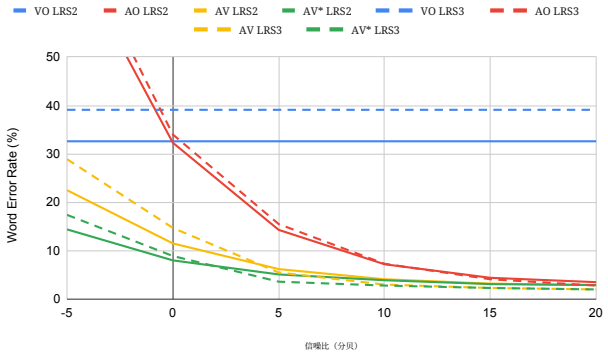
3.5. 噪声鲁棒性

我们使用各种类型的噪声测量模型的噪声鲁棒性，并将我们的音视高效 Conformer 与近期发表的方法进行比较。图6展示了来自 NoiseX语料库的白噪声和多人交谈噪声下，仅音频（音频输入）、仅视觉（视觉输入）和音视频模型在多个信噪比下的词错误率变化 [43]。我们发现，同时处理音频和视觉模态能够显著提升语音识别对多人交谈噪声的鲁棒性。此外，我们还按照先前工作的做法，在训练期间加入多人交谈噪声 [36, 29] ，并发现这能进一步提升模型在测试时的噪声鲁棒性。

对各种类型噪声的鲁棒性。 我们收集了各种录制音频噪声，包括声音和音乐。表10显示，在存在各种类型噪声的情况下，音频-视觉高效 Conformer 的表现始终优于其仅音频对应模型。这证实了我们的假设：当输入中存在音频噪声时，视听模型能够利用视觉模态辅助语音识别。



(a) 多人交谈噪声



(b) 白噪声

图6：**LRS2 和 LRS3 测试词错误率（%）随信噪比（分贝）的变化**。* 表示使用多人交谈噪声进行训练的实验。我们通过在多人交谈噪声和白噪声环境下评估模型，来衡量噪声鲁棒性。

表10：LRS3 测试词错误率（%）随信噪比（分贝）的变化。

噪声	Mode	信噪比（分贝） -5	0	5	10	15	20
多人说话噪声	AO	75.9	32.4	9.3	4.1	2.7	2.3
	AV	33.5	14.8	5.4	3.0	2.3	2.0
	AV*	11.2	4.9	3.1	2.5	2.2	2.0
白色	AO	77.6	34.0	15.5	7.3	4.1	2.8
	AV	28.9	14.7	5.5	3.0	2.3	2.0
	AV*	17.4	8.9	3.6	2.8	2.3	2.0
鸟类	AO	51.8	23.9	10.9	5.9	3.7	2.8
	AV	21.6	11.5	6.2	4.1	2.9	2.4
	AV*	15.9	8.3	4.9	3.4	2.7	2.4
电锯	AO	82.9	41.2	14.8	5.5	3.7	2.7
	AV	37.8	17.3	7.6	3.9	2.6	2.3
	AV*	25.8	10.8	5.0	3.2	2.4	2.3
jazz	AO	25.3	9.7	4.1	3.1	2.6	2.3
	AV	13.9	6.0	3.2	2.4	2.3	2.0
	AV*	10.6	4.2	2.8	2.4	2.2	2.0
街道下雨	AO	58.4	23.8	8.9	4.6	3.0	2.5
	AV	27.12	10.8	5.7	3.1	2.7	2.3
	AV*	15.9	6.9	3.8	2.7	2.3	2.2
洗碗	AO	47.8	24.5	11.5	6.0	3.7	2.8
	AV	21.3	11.5	6.1	3.6	2.8	2.3
	AV*	14.2	7.3	4.3	2.2	2.6	2.3
训练	AO	51.3	18.6	7.0	4.0	2.9	2.5
	AV	23.1	10.1	4.7	3.0	2.4	2.2
	AV*	14.5	6.2	3.5	2.6	2.3	2.2

与其他方法的比较。 我们将我们的方法与 Ma 等人 [29] 和 Petridis 等人 [36] 在 LRS2 测试集上给出的结果进行了比较。表11显示，在存在多人交谈噪声时，我们的视听模型取得了更低的词错误率，在 -5 分贝信噪比下词错误率为 9.7%，而 Ma等人 [29]为 16.3%。

表 11：与 Ma 等人 [29] 的比较。使用多人说话噪声时，LRS2 测试词错误率（%）随信噪比（分贝）的变化。

方法	Mode	信噪比（分贝） -5	0	5	10	15	20
Ma等人 [29]	VO	37.9	37.9	37.9	37.9	37.9	37.9
	AO*	28.8	9.8	7	5.2	4.5	4.2
	AV*	16.3	7.5	6.1	4.7	4.4	4.2
Ours	VO	32.6	32.6	32.6	32.6	32.6	32.6
	AO	70.5	27	8.6	4.7	3.4	3.1
	AV	25	11.2	5.1	3.2	2.8	2.6
	AV*	**9.7**	**5**	**3.4**	**2.9**	**2.8**	**2.6**

表 12：与 Petridis 等人 [36] 的比较。使用白噪声时，LRS2 测试词错误率（%）随信噪比（分贝）的变化。

方法	Mode	信噪比（分贝） -5	0	5	10	15	20
Petridis等人 [36]	VO	63.5	63.5	63.5	63.5	63.5	63.5
	AO*	85.0	45.4	19.6	11.7	9.4	8.4
	AV*	55.0	26.1	13.2	9.4	8.0	7.3
Ours	VO	32.6	32.6	32.6	32.6	32.6	32.6
	AO	73.1	32.3	14.3	7.2	4.4	3.5
	AV	22.5	11.5	6.2	4.1	3.2	2.9
	AV*	**14.4**	**8.0**	**5.1**	**3.9**	**3.1**	**2.9**

4. 结论

本文提出通过处理音频和视觉两种模态，提升近期提出的基于 Efficient Conformer CTC 的架构的噪声鲁棒性。我们证明，在模块之间加入多尺度 CTC 损失有助于提升识别性能，达到与最新自回归唇读方法相当的结果。我们还提出了块注意力，这是一种更简单、更高效的注意力机制，可替代第一个音频编码器阶段中的分组注意力。我们的音视高效 Conformer 在 LRS2 和 LRS3 测试集上分别取得了 2.3% 和 1.8% 的最先进性能。未来，我们希望探索其他技术，进一步提升模型的噪声鲁棒性，并缩小与近期唇读方法之间的差距。这包括在训练期间加入各种音频噪声，并利用预训练模型进行跨模态蒸馏。我们还希望在不损害识别性能的前提下降低视觉前端网络的复杂度，并尝试将 RNN-Transducer 学习目标用于流式应用。

致谢

本研究部分由亚历山大·冯·洪堡基金会（AvH）资助。

参考文献

[1] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman。深度视听语音识别。*IEEE* 模式分析与机器智能汇刊，2018 年。[2] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisser- man。Lrs3-ted：用于视觉语音识别的大规模数据集。*arXiv* 预印本 *arXiv:1809.00496*，2018 年。[3] Triantafyllos Afouras、Joon Son Chung 和 Andrew Zisser- man。ASR 只需你所需：用于唇读的跨模态蒸馏。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 2143–2147 页。IEEE，2020 年。[4] Yannis M Assael、Brendan Shillingford、Shimon Whiteson 和 Nando De Freitas。Lipnet：端到端句子级唇读。*arXiv* 预印本 *arXiv:1611.01599*，2016 年。[5] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Sub- biah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakan- tan、Pranav Shyam、Girish Sastry、Amanda Askell 等。语言模型是少样本学习者。神经信息处理系统进展，33:1877–1901，2020 年。[6] Adrian Bulat 和 Georgios Tzimiropoulos。我们距离解决 2D 和 3D 人脸对齐问题还有多远？（以及一个包含 230,000 个 3D 面部关键点的数据集）。载于*IEEE*国际计算机视觉会议论文集，第 1021–1030 页，2017 年。[7] Maxime Burchi 和 Valentin Vielzeuf。高效 Conformer：用于自动语音识别的渐进式下采样与分组注意力。载于*2021 IEEE*自动语音识别与理解研讨会（*ASRU*），第 8–15 页。IEEE，2021 年。[8] Joon Son Chung 和 Andrew Zisserman。野外唇读。载于亚洲计算机视觉会议，第 87–103 页。Springer，2016 年。[9] Joon Son Chung 和 AP Zisserman。侧面唇读。2017 年。[10] Ronan Collobert、Christian Puhrsch 和 Gabriel Synnaeve。Wav2letter：一种基于卷积网络的端到端语音识别系统。*arXiv* 预印本 *arXiv:1609.03193*，2016 年。[11] Jiankang Deng、Jia Guo、Evangelos Ververas、Irene Kot- sia 和 Stefanos Zafeiriou。Retinaface：用于野外环境的单次多级人脸定位。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第 5203–5212 页，2020 年。[12] Linhao Dong、Shuang Xu 和 Bo Xu。Speech-transformer：一种用于语音识别的无循环序列到序列模型。载于*2018 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第 5884–5888 页。IEEE，2018 年。[13] Haoqi Fan、Bo Xiong、Karttikeya Mangalam、Yanghao Li、Zhicheng Yan、Jitendra Malik 和 Christoph Feichten- hofer。多尺度视觉变换器。载于*IEEE/CVF*国际计算机视觉会议论文集，第 6824–6835 页，2021 年。[14] Alex Graves、Santiago Fern´andez、Faustino Gomez 和 J¨urgen Schmidhuber。连接时序分类：使用循环神经网络为未分段的序列数据标注。在*ICML*，第 369–376页，2006年。[15] Alex Graves、Abdel-rahman Mohamed 和 Geoffrey Hin- ton。使用深度循环神经网络进行语音识别。在*2013 IEEE*国际声学、语音与信号处理会议，第 6645–6649页。Ieee，2013年。[16] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Par- mar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zheng- dong Zhang、Yonghui Wu 等。Conformer：用于语音识别的卷积增强型 Transformer。*arXiv*预印本 *arXiv:2005.08100*，2020年。[17] Pengcheng Guo、Florian Boyer、Xuankai Chang、Tomoki Hayashi、Yosuke Higuchi、Hirofumi Inaguma、Naoyuki Kamo、Chenda Li、Daniel Garcia-Romero、Jiatong Shi 等。Conformer 助力 ESPnet 工具包的最新进展。在*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第5874–5878页。IEEE，2021年。[18] Wei Han、Zhengdong Zhang、Yu Zhang、Jiahui Yu、Chung-Cheng Chiu、James Qin、Anmol Gulati、Ruoming Pang 和 Yonghui Wu。ContextNet：利用全局上下文改进用于自动语音识别的卷积神经网络。*arXiv*预印本 *arXiv:2005.03191*，2020年。[19] Awni Hannun、Carl Case、Jared Casper、Bryan Catanzaro、Greg Diamos、Erich Elsen、Ryan Prenger、Sanjeev Satheesh、Shubho Sengupta、Adam Coates 等。Deep Speech：扩展端到端语音识别。*arXiv*预印本 *arXiv:1412.5567*，2014 年。[20] Kaiming He 、 Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。用于图像识别的深度残差学习。在*IEEE*计算机视觉与模式识别会议论文集，第770–778 页，2016年。[21] Kenneth Heafield。KenLM：更快速、更精简的语言模型查询。在第六届统计机器翻译研讨会论文集，第 187–197页，2011年。[22] Pavel Izmailov、Dmitrii Podoprikhin、Timur Garipov、Dmitry Vetrov 和 Andre w Gordon Wilson。权重平均可带来更宽的最优点和更好的泛化。*arXiv*预印本 *arXiv:1803.05407*，2018年。[23] Shigeki Karita、Nanxin Chen、Tomoki Hayashi、Takaaki Hori、Hirofumi Inaguma、Ziyan Jiang、Masao Someki、Nelson Enrique Yalta Soplin、Ryuichi Yamamoto、Xiaofei Wang 等。Transformer 与循环神经网络在语音应用中的比较研究。在*2019*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第449–456页。IEEE，2019年。[24] Diederik P Kingma 和 Jimmy Ba。Adam：一种随机优化方法。*arXiv*预印本 *arXiv:1412 6980*，2014年。。[25] Samuel Kriman、Stanislav Beliaev、Boris Ginsburg、Joce- lyn Huang、Oleksii Kuchaiev、Vitaly Lavrukhin、Ryan Leary、Jason Li 和 Yang Zhang。QuartzNet：采用一维时间-通道可分离卷积的深度自动语音识别。在*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第6124–6128页。IEEE，2020年。[26] Taku Kudo 和 John Richardson。SentencePiece：一种用于神经文本处理的简单、语言无关的子词分词器和反分词器。在*EMNLP*，第66 71页，2018 – 。

用于神经文本处理。在*EMNLP*，第66 71页，2018 – 。[27] Jaesong Lee 和 Shinji Watanabe。用于基于 CTC 的语音识别的中间损失正则化。在*ICASSP 2021- 2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第6224–6228页。IEEE，2021年。[28] Jason Li、Vitaly Lavrukhin、Boris Ginsburg、Ryan Leary、Oleksii Kuchaiev、Jonathan M Cohen、Huyen Nguyen 和 Ravi Teja Gadde。Jasper：端到端卷积声学模型。*arXiv*预印本 *arXiv:1904.03288*，2019年。[29] Pingchuan Ma、Stavros Petridis 和 Maja Pantic。使用 Conformer 的端到端音频-视觉语音识别。在*ICASSP 2021-2021 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第7613–7617页。IEEE，2021年。[30] Pingchuan Ma、Stavros Petridis 和 Maja Pantic。真实环境中的多语言视觉语音识别。*arXiv*预印本 *arXiv:2202.13084*，2022年。[31] Somshubra Majumdar、Jagadeesh Balam、Oleksii Hrinchuk、Vitaly Lavrukhin、Vahid Noroozi 和 Boris Ginsburg。Citrinet：缩小非自回归与自回归端到端自动语音识别模型之间的差距。*arXiv*预印本 *arXiv:2104.01721*，2021年。[32] Takaki Makino、Hank Liao、Yannis Assael、Brendan Shillingford、Basilio Garcia、Otavio Braga 和 Olivier Sio- han。用于音频-视觉语音识别的循环神经网络转导器。在*2019*年*IEEE*自动语音识别与理解研讨会（*ASRU*），第905–912页。IEEE，2019年。[33] Jumon Nozaki 和 Tatsuya Komatsu。通过以中间预测为条件，放宽基于 CTC 的自动语音识别中的条件独立假设。*arXiv*预印本 *arXiv:2104.02724*，2021年。[34] Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。在*2015*年*IEEE*国际声学、语音与信号处理会议（*ICASSP*），第5206–5210页。IEEE，2015年。[35] Daniel S Park、Yu Zhang、Chung-Cheng Chiu、Youzheng Chen、Bo Li、William Chan、Quoc V Le 和 Yonghui Wu。大型数据集上的 SpecAugment。在*ICASSP*，第6879–6883页，2020年。[36] Stavros Petridis、Themos Stafylakis、Pingchuan Ma、Geor- gios Tzimiropoulos 和 Maja Pantic。采用 CTC/注意力混合架构的音频-视觉语音识别。在*2018*年*IEEE* 口语技术研讨会（*SLT*），第513–520页。IEEE，2018年。[37] KR Prajwal、Triantafyllos Afouras 和 Andrew Zisserman。结合视觉注意力的子词级唇读。在*IEEE/CVF*计算机视觉与模式识别会议论文集，第5162–5172页，2022年。[38] Prajit Ramachandran、Barret Zoph 和 Quoc V Le。搜索激活函数。*arXiv*预印本 *arXiv:1710.05941*，2017年。[39] Dmitriy Serdyuk、Otavio Braga 和 Olivier Siohan。音频-视觉语音识别值得使用32x32x8体素。在 *2021*年 *IEEE* 自动语音识别与理解研讨会 *(ASRU)*，第796–802页。IEEE，2021年。[40] Brendan Shillingford、Yannis Assael、Matthew W Hoff- man、Thomas Paine、C´ıan Hughes、Utsav Prabhu、Hank Liao、Hasim Sak、Kanishka Rao、Lorrayne Bennett 等。大规模视觉语音识别。*arXiv* 预印本 *arXiv:1807.05162*，2018年。[41] Joon Son Chung、Andrew Senior、Oriol Vinyals 和 Andrew Zisserman。在野外进行句子唇读。载于*IEEE*计算机视觉与模式识别会议论文集，第6447–6456页，2017年。[42] George Sterpu、Christian Saam 和 Naomi Harte。基于注意力的音视频融合用于稳健的自动语音识别。载于第*20*届 *ACM* 多模态交互国际会议论文集，第111–115页，2018年。[43] Andrew Varga 和 Herman JM Steeneken。自动语音识别评估：II。NOISEX-92：用于研究加性噪声对语音识别系统影响的数据库与实验。《语音通信》，12(3)：247–251，1993年。[44] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszko- reit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。注意力机制即所需的一切。神经信息处理系统进展，30，2017年。[45] Bo Xu、Cheng Lu、Yandong Guo 和 Jacob Wang。判别式多模态语音识别。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第14433–14442页，2020年。[46] Jianwei Yu、Shi-Xiong Zhang、Jian Wu、Shahram Ghorbani、Bo Wu、Shiyin Kang、Shansong Liu、Xunying Liu、Helen Meng 和 Dong Yu。针对 LRS2数据集的重叠语音音视频识别。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第6984–6988页。IEEE，2020年。[47] Qian Zhang、Han Lu、Hasim Sak、Anshuman Tripathi、Erik McDermott、Stephen Koo 和 Shankar Kumar。Transformer传感器：一种采用 Transformer编码器和 RNN-T损失的流式语音识别模型。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议 *(ICASSP)*，第7829–7833页。IEEE，2020年。[48] Xingxuan Zhang、Feng Cheng 和 Shilin Wang。基于时空融合的卷积序列学习用于唇读。载于*IEEE/CVF*国际计算机视觉会议论文集，第713–722页，2019年。[49] Ya Zhao、Rui Xu、Xinchao Wang、Peng Hou、Haihong Tang 和 Mingli Song。聆听嘴唇：通过蒸馏语音识别器改进唇读。载于*AAAI*人工智能会议论文集，第34卷，第6917–6924页，2020年。