上海交通大學

SHANGHAI JIAO TONG UNIVERSITY

# 学士学位论文

BACHELOR'S THESIS

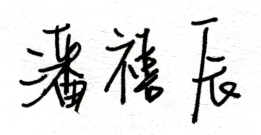
论文题目：基于预训练模型的多模语音识别系统


| | |
|---|---|
| 学生姓名： | 潘禧辰 |
| 学生学号： | 518021910497 |
| 专　　业： | 计算机科学与技术 |
| 指导教师： | 林洲汉 |
| 学 院(系)： | 电子信息与电气工程学院 |

# 上海交通大学

# 学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名： 

日期：2022 年 5 月 31 日

# 上海交通大学

# 学位论文使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

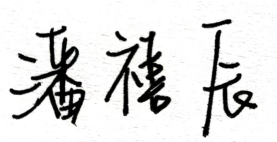
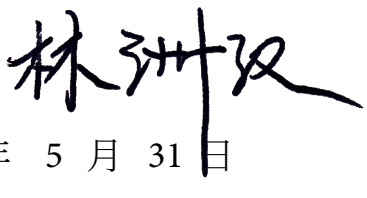
本学位论文属于 ☑**公开论文**

☐**内部论文**，☐ 1 年/☐ 2 年/☐ 3 年 解密后适用本授权书。

☐**秘密论文**，____ 年（不超过 10 年）解密后适用本授权书。

☐**机密论文**，____ 年（不超过 20 年）解密后适用本授权书。

（请在以上方框内打“✓”）

学位论文作者签名：  指导教师签名： 

日期：2022 年 5 月 31 日　　日期：2022 年 5 月 31 日

# 基于预训练模型的多模语音识别系统

## 摘　要

训练基于注意力机制的模型需要大量数据，而获得标记的对齐多模态数据成本高昂，特别是对于音视频语音识别任务。因此，利用未标记的单模态数据是非常有意义的。另一方面，尽管大规模自监督学习的有效性在音频和视觉模态中都得到了证实，但如何在视频语音识别任务中应用预训练模型，以及如何将这些预训练模型整合到多模场景中仍然需要探索。对于第一个问题，在这项工作中我们通过使用随机初始化的三维卷积替换通用视觉自监督预训练模型中的二维卷积，使其能够被用于有效提升视频语音识别性能。在词级视频语音识别上，我们使用预训练视觉前端，通过注意力遮罩来整合词边界信息，及引入视素级 CTC 损失辅助训练，提出的模型在 LRW 数据集上达到了 89.1% 的准确率，将最佳性能绝对提升了 0.6%。对于第二个问题，我们将预训练过的音频和视觉前端的部分参数整合到一个句级多模音视频语音识别框架中，该框架通过结合 CTC 和 seq2seq 解码识别对齐的音视频输入。实验表明，从单模自监督学习中继承下来的两个前端合作良好，使得多模框架可以通过微调产生有竞争力的结果。即使没有外部语言模型，提出的模型也大幅提高了 LRS2 数据集上句级音视频语音识别任务的表现，达到了 2.6% 的错词率，相对目前最优模型提高了 30%。

**关键词：**多模态深度学习，语音识别，自监督学习，音视频语音识别，唇读

# MULTIMODAL AUDIO-VISUAL SPEECH RECOGNITION SYSTEM BASED ON PRE-TRAINED MODELS

## ABSTRACT

Training attention mechanisms based models requires a large amount of data and obtaining labeled aligned multimodal data is costly, especially for audio-visual speech recognition. Thus it makes much sense to utilize unlabeled unimodal data. On the other side, although the effectiveness of large-scale self-supervised learning has been well established in both audio and visual modalities, how to apply pre-trained models in video speech recognition and how to integrate these pre-trained models into multimodal scenarios remain underexplored. For the first problem, in this work, we replace the 2-D convolution in the generic visual self-supervised pre-training model by using randomly initialized 3-D convolution so that it can be used to improve the performance of video speech recognition effectively. For word-level video speech recognition, we use a above-mentioned pre-trained visual front-end, introduce attention masks to integrate word boundary information and a viseme-level CTC loss to aid training. The proposed model achieves 89.1% accuracy on the LRW dataset, with an absolute improvement of 0.6% over the state-of-the-art performance. For the second problem, we integrate partial parameters of the pre-trained audio and visual front-ends into a sentence-level multimodal audio-visual speech recognition framework that recognizes aligned audio-visual inputs by using a combination of CTC and seq2seq decoding. Experiments show that the two front-ends inherited from unimodal self-supervised learning cooperate well, allowing the multimodal framework to obtain competitive results through fine-tuning. Even without an external language model, the proposed model improves the state-of-the-art performance of the sentence-level audio-visual speech recognition on the LRS2 dataset by a large margin, achieving a 2.6% word error rate, with a relative improvement of 30%.

**Key words:** multimodal deep learning, speech recognition, self-supervised learning, audio-visual speech recognition, lipreading

# 目　录

# 插图索引

# 表格索引

# 算法索引

# 第一章　绪论

本文中我们将介绍两部分工作，第一部分是词级视频语音识别（word-level visual speech recognition）；第二部分是句级多模音视频语音识别（sentence-level audio-visual speech recognition）。这两个任务是目前多模音视频语音识别（multimodal audio-visual speech recognition）领域最受关注的两个任务。本文的组织结构如下，首先我们将在第二章讲述多模音视频语音识别领域的相关工作，包括自动语音识别、词级视频语音识别、多模音视频语音识别和自监督学习。然后，我们将在第三章详细介绍词级视频语音识别的方法及其相应的实验。这将包括我们如何使用 MoCo v2 前端，如何使用注意力遮罩整合词边界信息，以及视素级 CTC 辅助损失的定义，之后是此部分的实验结果。第四章是关于句级音视频语音识别模型，我们将介绍其使用的两个自监督单模特征提取器，模型的实现细节，以及详细的实验结果。第五章将介绍我们搭建的音视频语音识别工具链，包括其设计框架及功能。最后，我们将给出整个论文的结论，总结我们方法的使用，并谈及一些未来方向。

## 1.1　词级视频语音识别介绍

视频语音识别（visual speech recognition），也被称为唇读（lip reading），依靠无声的视频（silent video）来识别说话者说出的内容。因为在噪声环境中来自视觉模态的信息可以极大地提高纯音频（audio only）自动语音识别（auto speech recognition）的性能[1]，这项任务具有很高的研究价值并吸引了大量的研究关注。

词级视频语音识别任务也被称为词级唇读（word-level lip reading），本质上是序列分类任务，具体而言是将一段包含某个词的纯视觉，或称无声视频（slient video）输入进行分类。目前最大的公开英语词级视频语音识别数据集是 LRW[2]（Lip Reading in the Wild），该数据集中视频片段（video clip）固定 29 帧长，由 500 个类组成，每类视频对应一个目标单词，目标单词处于视频片段的正中央，同时由于数据集是从 BBC 节目中收集的，视频片段中除目标词之外，还包含邻近的其他单词。数据集提供词边界（word boundary）用于说明哪些帧是目标词所在的帧。

关于如何有效的使用词边界信息，现有工作做了很多探索，Stafylakis et al.[3], Feng et al.[4]提出在循环神经网络（recurrent neural network）中将词边界加入到特征中，使用 0-1 来表示该帧是否位于词边界内，获得了很大的提升。原因是循环神经网络有强大的门控机制，可以利用词边界的额外特征。此外，循环神经网络能够对长距离的时间依赖性进行建模，在词边界外的帧能够提供关于此特定视频片段的特征（如说话人、姿势和其他视频片段级特征）。因此词边界外的帧可以被循环神经网络通过在单元（unit）中积累相关信息进行利用。此外，边界外的帧还带有目标词的上下文，有助于目标词的识别。

因为 LRW 数据集的序列长度较短，只有 29 帧，基于卷积神经网络（convolutional neural network）的词级视频语音识别模型[5]在不使用词边界信息的情况下，也能够达到与循环神经网络同样的表现。而在序列长度更长的中文词级视频语音识别数据集 LRW-1000[6]上，卷积神经网络的表现距离最好的基于循环神经网络的模型[4]仍然有较大差距，且在基于卷积神经网络的模型中结合词边界信息的方法仍然没有得到充分探索。在 LRW 数据集上，现有最优的基于注意力机制（attention mechanism）的模型 Wiriyathammabhum[7]性能距离上述两种

模型仍有较大差距，原因是基于注意力机制的模型对输入位置不敏感，在短序列、小数据量的任务上不如有对位置有较强先验的循环神经网络和卷积神经网络模型，更适合长序列、大数据量的任务。此外，同基于卷积神经网络的模型一样，基于注意力机制的模型也无法有效地利用词边界信息。

在本文第三章中，我们提出了使用注意力遮罩（attention mask）的方法来将词边界整合到 Transformer 模型中，同时使用基于 MoCo v2[8]的预训练视觉前端，使用该视觉前端的动机将在1.2节中介绍。提出的模型在 LRW 数据集上达到了 88.8% 的准确率，是目前该数据集的最优模型（state-of-the-art model）。在此基础上我们加入视素级别 CTC（viseme-level connectionist temporal classification）损失辅助学习，能够进一步提升到 89.1% 的准确率，相比目前最优模型 Kim et al.[9]绝对提升了 0.6%。

## 1.2 句级多模音视频语音识别介绍

多模音视频语音识别是一项同时利用人声的音频输入和唇部动作的视觉输入进行语音识别的任务。近年来，它一直是多模态成功应用的领域之一。

由于有标注且对齐的多模数据量有限，而且从视觉输入进行识别，即视频语音识别的难度很大，所以这是一项具有挑战性的任务。现有的句级音视频语音识别模型倾向于使用额外的数据来提高系统的性能，其形式是在训练过程中加入一个预先的监督学习（supervised learning）阶段。例如，在句级音视频语音识别任务的学习之前，许多现有的方法依靠额外的词级视频语音识别任务来引导模型对视觉特征的学习。Petridis et al.[10], Zhang et al.[11]在 LRW 数据集上训练他们的视觉前端。Afouras et al.[1], Afouras et al.[12]将 MV-LRS 数据[13]切割成含有单个单词的视频片段，通过分类对模型进行预训练，Afouras et al.[14]中也使用相同方法将 VoxCeleb 数据集[15]用于预训练。然而即使有这些额外的监督学习任务，学习一个有效的视觉前端仍然是十分困难的。有时需要进行课程学习（curriculum learning），以使学到的视觉前端适应句级音视频语音识别任务[1]，大规模音视频语音识别的端到端学习直到最近才获得成功[16]。另一方面，目前流行的音视频语音识别模型[1, 16]都使用基于注意力机制的时序建模模块，训练这种模型对训练数据的规模提出了更高的要求。

有标注且对齐的多模数据量有限，然而目前在音频和视觉两个模态已经分别有大量的单模无标注数据。详细对比我们发现，目前有标注的对齐多模音视频语音识别数据集主要为 LRW[2]（157 小时）、LRS2[1]（224 小时）、LRS3[12]（438 小时），共计 819 小时数据。作为对比，目前音频中我们可以使用无标注的 Libri-Light[17]数据集，其中包含 60K 小时的音频数据，规模远大于有标注的对齐多模数据；视觉模态中有 ImageNet[18]（14M 张图片）、COCO[19]（123K 张图片）等大量分类和检测等任务的数据集可供使用。

直观的想法是利用自监督学习来使用这些大规模无标注单模数据提升多模音视频语音识别的性能，单模的自监督学习已经被公认为是一种从无标注数据中学习一般表征的范式，例如在自然语言处理中[20-21]，语音识别中[22]，以及计算机视觉中[23-25]。但在多模音视频语音任务上使用自监督学习的方法还没有得到充分的探索。Shukla et al.[26]是这方面的少数尝试之一，其提出从音频输入中预测唇部动作。他们提出的模型达到了突出的情感识别性能，但在语音识别方面相对较弱。

一个简单且直接的想法是使用单模自监督预训练模型作为多模音视频语音识别中音频和视觉两个模态的特征提取器。在音频模态中，wav2vec[27]、vq-wav2vec[28]、wav2vec 2.0[22]、WavLM[29]、HuBERT[30]等自监督预训练模型都展示了其在自动语音识别任务上的突出性能。

然而在视觉模态，目前还没有可用于视频语音识别的自监督预训练模型。但对于通用视觉表征提取器，MoCo[23]、SimCLR[24]、SimSiam[31]、BYOL[25]等模型都在分类和检测等任务上展示了其出色的性能。使用此类模型作为视觉模态的特征提取器存在一些困难。由于在音视频语音识别中，帧之间的唇部动作是非常重要的，且预训练使用的图像多与唇部无关，为针对单帧图像的任务定制的预训练的视觉模型是否能够适用于音视频语音识别仍然是未知的。

在这项任务中，我们使用在大规模的 Libri-Ligh 数据集上预训练的 wav2vec 2.0[22]作为我们的音频前端。对于视觉前端，我们使用在 ImageNet[18]数据集上预训练的 MoCo v2[8]，并用一个三维卷积代替 MoCo v2 中的第一个卷积层，然后通过在 LRW 数据集上的词级视频语音识别任务对其进行微调。总的来说，我们的方法不需要课程学习阶段，并且相较未使用预训练模型的方法训练时间也得到了降低。实验结果表明，我们的新前端在纯音频和纯视觉设置中都大大超过了以前的前端，在音视频设置中，我们取得了新的最先进表现，在 LRS2 数据集上达到了 2.6% 的错词率，相对目前最优模型提升 30%。据我们所知，这是第一个在音视频语音识别的多模设置中成功应用单模预训练模型的工作。

## 1.3 研究贡献总结

本文的贡献可以归纳为以下几点：

**成功使用通用视觉预训练模型提升视频语音识别性能：**我们通过将通用视觉自监督预训练模型中的二维卷积更换为随机初始化的三维卷积，使得预训练特征提取器能够捕获视觉帧之间的时间相关性，同时减轻预训练数据域间差异（domain shift）带来的影响。

**提出了在基于注意力机制的模型中使用注意力遮罩来整合词边界信息：**我们提出使用注意力遮罩来整合词边界信息，通过目标词所在帧整合词边界外帧的环境和上下文信息，通过特殊的 `[CLS]` 嵌入整合目标词所在帧的信息。使得基于注意力机制的模型也能够有效的利用词边界信息帮助词级视频语音识别任务的学习。

**提出了使用视素级 CTC 损失辅助词级视频语音识别的学习：**我们使用视素级 CTC 损失给予解码器输出额外的视素级监督信号，帮助目标词所在帧特征的学习。

**提出了应用单模预训练模型提升多模音视频语音识别性能的方法：**我们提出将在大规模单模数据集上训练过的音频和视觉前端作为特征提取器，整合到一个更大的多模音视频语音识别框架中的方法，极大提升了多模场景下的性能。

**编写了多模音视频语音识别工具链：**我们以本文中提出的两个模型的实现为基础，构建预处理和数据输入管道，整理预定义模块和实用工具，提供训练代码框架和模型复现用例，完成了实用的多模音视频语音识别工具链。

## 1.4 本章小结

本章介绍了本文的整体结构，词级视频语音识别和句级多模音视频语音识别两个任务的相关背景，本文的动机和主要工作内容。最后我们还对的研究贡献进行了总结。

# 第二章　相关工作

## 2.1　自动语音识别

自动语音识别通常指通过音频输入识别说话内容的任务，最早的隐马尔科夫模型——深度神经网络（hidden markov model-deep neural network ）自动语音识别系统广泛使用人工音频特征，如梅尔倒频谱系数（mel frequency cepstral coefficients）输入[32]。但新兴的声学模型（acoustic model）系统倾向于使用原始频谱特征[33]或原始音频（raw audio）作为输入，并使用可训练的卷积语音前端对其进行特征提取[34-35]。采取原始信号的端到端系统在语音前端可以达到相同的性能，相比人工音频特征降低了系统的复杂性。为了学习将音频表征识别为符号，端到端（end to end）的声学模型通常采用 CTC 损失[36]或 seq2seq（sequence to sequence）损失训练，Watanabe et al.[37]显示，CTC 和 seq2seq 的混合训练和解码可以进一步改善自动语音识别的性能。

## 2.2　词级视频语音识别

最早在大规模唇读数据集 LRW 上使用深度学习方法进行词级视频语音识别的工作 Chung et al.[2]使用了一个基于 VGG-M 的多塔（multi tower）卷积神经网络架构。现代词级视频语音识别模型通常由一个从唇部感兴趣区域（region of interest）提取特征的前端和一个用于时序建模的后端组成。Stafylakis et al.[38]首次引入三维卷积神经网络作为前端，其由一个三维卷积层接着一个深度二维卷积网络构成，因为其较好的时序特征提取能力，已被广泛用作前端[5, 16, 39]。对于后端，基于循环神经网络的模型，特别是 LSTM[40]（long short-term memory）网络和 GRU[41]（gated recurrent unit）网络因其较好的短序列建模能力以及突出的整合词边界信息的能力被广泛应用[4, 38, 42-44]。基于卷积神经网络的后端也得到了广泛应用，时序卷积网络（temporal convolution networks）被应用于 Ma et al.[5], Martínez et al.[39], Ma et al.[45]中，目前最优模型 Kim et al.[9]也使用了时序卷积网络作为后端。基于注意力机制的后端如 Transformer[46]在 Wiriyathammabhum[7]中得到尝试。

在 Martínez et al.[39]之前，由于循环神经网络难以直接进行训练，训练流程通常采用三阶段的方法[38, 42]。首先，用单层的时序卷积网络后端来对前端进行训练。然后，使用随机初始化的双向 GRU 或双向 LSTM 模块取代时序卷积网络后端，冻结前端参数单独对后端进行训练。最后，将整个网络一起进行微调。Martínez et al.[39]将训练过程简化为端到端的训练，这大大减少了训练时间。目前基于卷积神经网络的模型 Ma et al.[5] 与基于循环神经网络的模型 Feng et al.[4]都可以达到相同的准确率。最优模型 Kim et al.[9]通过加入一个多头视觉音频记忆模块（multi-head visual-audio memory）来捕获视觉和音频之间的关系，可以分辨出唇部动作相同但发音不同的单词。

## 2.3　多模音视频语音识别

最早的音视频语音识别工作可以追溯到大约二十年前，Dupont et al.[47]展示了手工制作的视觉特征可以改善基于隐马尔可夫模型的自动语音识别系统。第一个现代音视频语音识别系统是在 Afouras et al.[1]中提出的，其中使用了深度神经网络。此后，该领域一直在快速

发展。大多数工作都致力于架构的改进，例如，Zhang et al.[11]提出了时间焦点块（temporal focal block）和空间-时间融合策略，Lee et al.[48]探索了使用 Transformer 的跨模态关注机制。

另一条研究路线集中在使用更加多样化的学习方案以提高音视频语音识别的性能。Li et al.[49]使用了一个跨模态的师生训练方案。Paraskevopoulos et al.[50]提出了一个多任务学习方案，使模型在字符（character）和子词（sub-word）级别上都能识别。Shukla et al.[26]中探索了自监督学习的应用，其利用了跨模态设置，从音频输入中预测视频帧。

音视频语音识别系统的端到端学习首次出现在 Tao et al.[51]中，但是其使用的数据集比 LRS2 简单得多。最近的最优模型[16]通过使用 Conformer[52]声学模型和混合 CTC/注意力解码器使 LRS2 的端到端学习成为可能。

## 2.4 自监督学习

自监督学习（self-supervised learning）近年来一直十分流行，因为它能够通过不需要标注的简单任务来学习数据的一般表征。对比学习[53]（contrastive learning）已经成为这个领域中最具影响力的方案。在自然语言处理中，单向或双向的语言建模[20-21]（language modelling）已经被用于大幅提高各种任务的性能。在自动语音处理中，对比性预测编码[22]（contrastive predictive coding）已在语音识别任务中显示出突出作用。在视觉领域，早期的工作通过基于图像处理的方法创建了自监督的任务，如失真[54]（distortion），着色[55]（colorization）和上下文预测[56]（context prediction）。最近，使用对比学习进行通用视觉表征学习的模型取得了成功，如 MoCo[8, 23], SimCLR[24], BYOL[25]等，其在下游的分类和检测等任务上均具有突出表现。

## 2.5 本章小结

本章介绍了现有的自动语音识别、词级视频语音识别、多模音视频语音识别和自监督学习的工作。词级视频语音识别对应本文第三章相关内容，多模音视频语音识别对应本文四章相关内容。自动语音识别中的时序建模和解码部分的设计为多模音视频语音识别提供了有价值的方案。我们提出的词级视频语音识别模型和句级音视频语音识别模型都应用了自监督学习预训练模型作为特征提取前端。

# 第三章　词级视频语音识别

## 3.1　模型

本章节将对我们提出的词级视频语音识别模型进行介绍。模型主要分为前端，后端和分类器三部分，前端通过由 MoCo v2 初始化的卷积神经网络来对每个视频帧进行特征提取，后端进行时序建模，将所有视频帧经过前端提取的特征进行整合，再经过分类器完成分类。

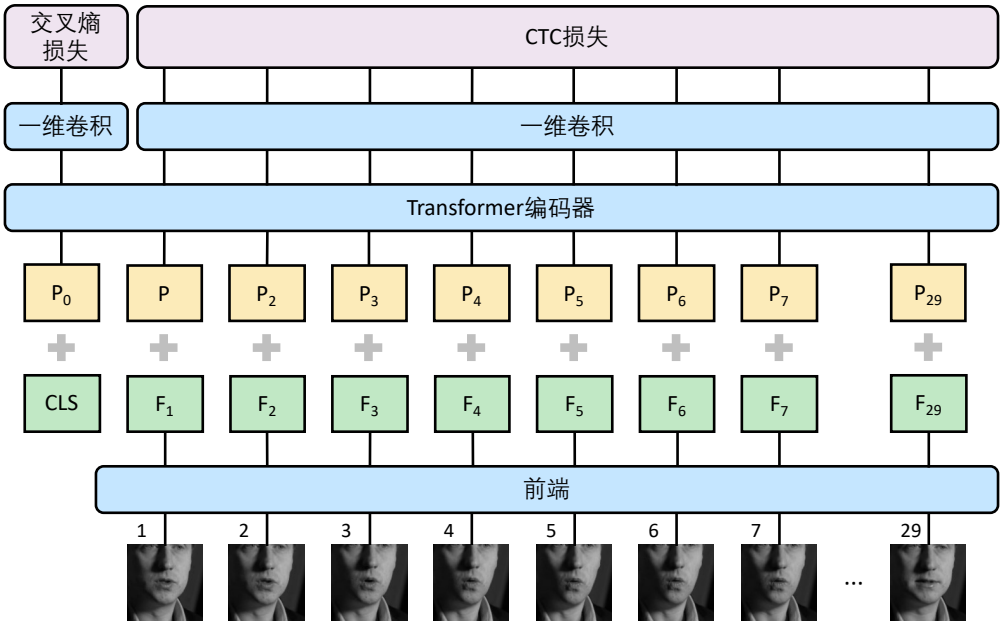



**图 3–1　词级视频语音识别模型**

**F 代表视频帧特征，CLS 代表特殊的可学习的 [CLS] 嵌入（embedding），P 代表位置编码**

图3–1中展示了词级视频语音识别模型的整体架构，模型的输入是 29 帧灰度图像，分别是每帧裁剪出来的感兴趣区域。经过前端进行处理之后得到每个视频帧的特征 $F$，我们额外加入一个可学习的 [CLS] 嵌入，将这些特征加上绝对位置编码 $P$，之后送到 Transformer 编码器（encoder）中进行时序建模。[CLS] 嵌入的输出向量经过一维卷积完成分类，通过交叉熵（cross-entropy）损失进行优化。视频帧的输出向量经过一维卷积得到 CTC 概率，通过 CTC 损失辅助进行优化。

表3–1中说明了各个部分中视频流特征维度的变化。模型输入是 $T \times 112^2 \times 1$ 大小的视频序列。在前端中，经过三维卷积之后，特征图（feature map）的宽和高降低到 28，通道数提升到 64。再经过 MoCo v2 中预训练的 ResBlock 继续提取特征，将每个视频帧提取出 2048 维的特征。在后端对提取出的特征先使用一维卷积进行降维到 $d_{model} = 512$ 维，之后通过 Transformer 编码器进行视频帧特征的整合，得到 $d_{model} = 512$ 维整个视频片段的特征。在分类器部分，使用一维卷积对该特征进行分类，得到 $N = 500$ 维的分类结果。

| 部分 | 模块 | 视频输入<br>$T \times 112^2 \times 1$ |
|---|---|---|
| 前端 | 三维卷积 | $T \times 28^2 \times 64$ |
| | MoCo v2 | $T \times 2048$ |
| 后端 | 一维卷积 | $T \times d_{model}$ |
| | Transformer 编码器 | $1 \times d_{model}$ |
| 分类器 | 一维卷积 | $N$ |

**表 3–1 视频流的特征维度**

**{时间维大小 × (空间维大小$^2$) × 通道数} 代表特征维度，$T = 29$ 代表视频帧的数量，$d_{model} = 512$ 代表模型维度，$N = 500$ 代表类别数。**

### 3.1.1 前端

前端是用于捕捉唇部动作并在其输出表示中反映唇部形态及其差异的模块，前端能够提取每一视频帧的特征。

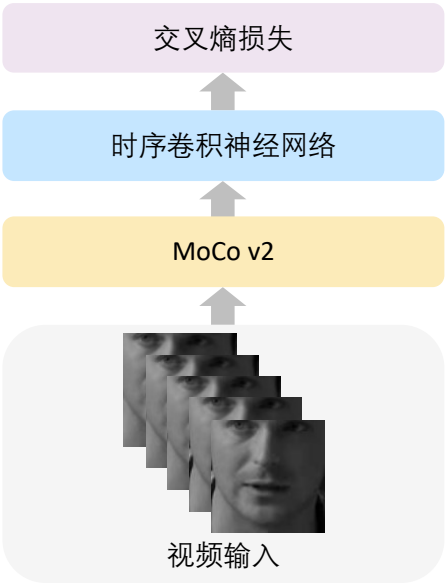
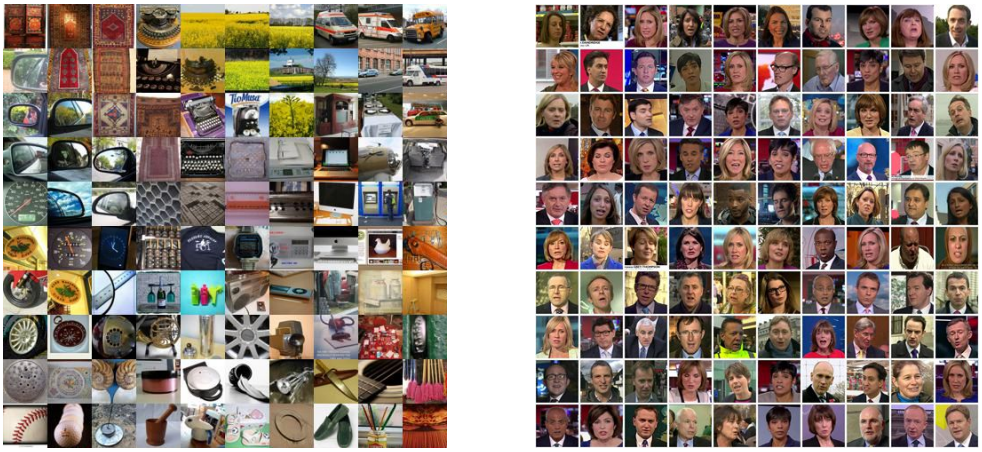



**图 3–2 直接使用 MoCo v2 作为前端的词级视频语音识别模型**

一个在前端应用预训练模型的简单想法是直接用其提取原始视频帧的特征。我们首先按照 Stafylakis et al.[38]中的模型搭建了一个简单的基线模型，将视频输入经过预训练过的 MoCo v2（ResNet-50 架构），得到每个视频帧的特征，之后我们使用了两个时间卷积层，每个卷积层后面都有批标准化[57]（batch normalization）、ReLU 和最大池化层（max pooling），这两个时间卷积层将时间维降低了 2 倍。之后，我们使用一个平均池化层（average pooling）来将特征整合成一个向量，通过一个线性层完成分类。分类结果通过交叉熵损失进行优化。

经过实验我们发现这个模型的效果欠佳，原因主要有两点，其一是 MoCo v2 预训练与词级视频语音识别所用数据集之间的域间差异（domain shift）；其二是 MoCo v2 缺少对帧间时间联系的建模。

图3–3展示了 MoCo v2 预训练所用的 ImageNet 数据集与词级视频语音识别所用的 LRW

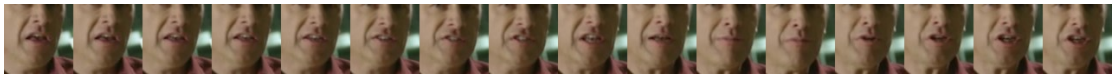


a) ImageNet[18]中的图像　　b) LRW[2]中的图像

**图 3–3 ImageNet 与 LRW 的域间差异**
**ImageNet 中图像为各种物体，LRW 中图像为说话人的脸，二者存在域间差异**

数据集之间的域间差异。由于在视频语音识别任务中同一视频片段中的帧在内容上基本相似，都是说话人的面部或唇部图像，而 MoCo v2 通过预训练任务学习到的是反映整个图像内容的一般表征，这将导致视频片段中所有帧的输出类似，而帧中对于视频语音识别任务非常重要的唇部形态信息则被忽略了。

a) ImageNet 中鱼的图像



b) LRW 中视频片段的图像

**图 3–4 ImageNet 相比 LRW 时序信息的缺失**
**ImageNet 中图像为各个独立的场景，无时序信息；LRW 中图像为一段连续的说话视频**

MoCo v2 模型的预训练任务是拉进同一张图片经过两种不同数据增强方式进行增强，并由编码器编码得到的特征，同时拉远不同图片之间的特征。这种预训练任务在不同图像间进行训练，各个图像都是无时序信息的独立场景。而视频语音识别任务关注的除了单帧图像能够提供的唇部形态，还有帧之间的唇部动作。图3–4展示了 ImageNet 数据集相比 LRW 数据集的时序信息缺失。故 MoCo v2 缺少对帧间的时间联系进行建模的能力，也导致了其直接用在视频语音识别任务上的表现欠佳。

为了解决直接应用 MoCo v2 模型产生的域间差异，以及帧间的时间联系建模的缺失。如图3–5所示，我们仅保留 MoCo v2 中后四个 ResBlock，将前边的预训练二维卷积替换为三维卷积。三维卷积包括一个具有 64 个通道的 $5 \times 7 \times 7$(时间 × 宽度 × 高度) 卷积核大小的三维卷积层，然后是批量归一化和 ReLU。提取的特征图之后通过一个时空最大池化层来降低空间和时间上的尺寸，最终输出的特征图大小为 $T \times 28^2 \times 64$，该尺寸与 MoCo v2 的第一个 ResBlock 的输入相同，提供了一个兼容的接口将 MoCo v2 的后四个预训练 ResBlock 参数转

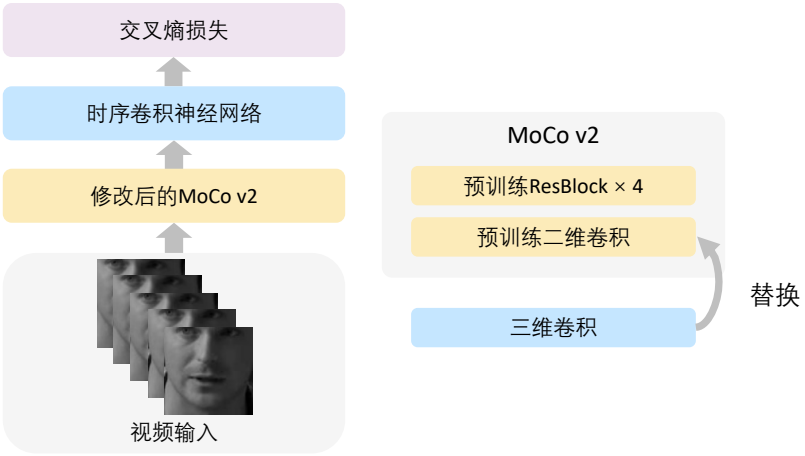



**图 3–5 使用调整后的 MoCo v2 作为前端的词级视频语音识别模型**

移到视频语音识别任务中。同时，我们将 RGB 视频图像先转换为灰度再送入模型，这可以有效防止模型学习色差信息。如此修改的好处首先是三维卷积能够在视觉前端捕捉唇部区域的动态信息；其次是引入了新的随机初始化模块，通过视频语音识别的监督任务使得预训练前端可以重新学习合理的唇部形态和动作的表征。经过实验，发现修改后的前端能够有效提升视频语音识别性能，详细结果记录在3.2节。

### 3.1.2 后端

后端是进行时序建模的模块，其能够学习前端提取出来的视频帧的特征之间的时序联系，将其这些特征整合为一个向量表示，便于分类器进行词级分类。

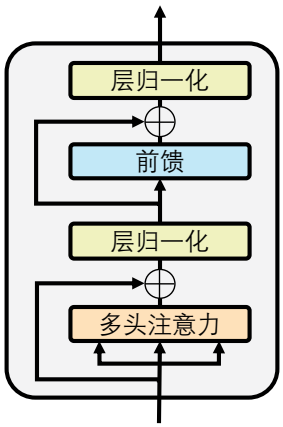



**图 3–6 Transformer 编码器层**

我们使用了一维卷积和 6 层 Transformer 编码器[46]作为后端。一维卷积包括一个输入通道数为 2048，输出通道数为 $d_{model}$，步长为 1，卷积核大小为 1 的一维卷积层，以及批量归一化和 ReLU，其目的在于对前端提取的特征进行降采样。每层 Transformer 编码器层的结构如图3–6所示，包含一个多头自注意层（multi-head self-attention），输入张量同时作为注意力的查询（query），键（key）和值（value），每个注意力头（attention head）都进行缩放点积注意力（scaled dot-product attention）计算，其输入包括维度为 $d_k$ 的查询和键，以及维度为

$d_v$ 的值，注意在 Transformer 编码器中 $d_k = d_v$

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \tag{3–1}$$

其中 $Q$ 代表查询，$K$ 代表键，$V$ 代表值。

多头注意力是指在计算缩放点积注意力时，使用不同的可学习的线性投影将查询、键和值分别投影到 $d_k$、$d_k$ 和 $d_v$ 维度上，在这些投射的查询、键和值的每个版本上并行执行缩放点积注意力计算，产生 $d_v$ 维的输出值。这些值按维度进行拼接后再次进行投影，从而得到最终的值。多头注意力允许模型关注来自不同位置上不同表征子空间的信息。

$$\begin{aligned} \text{MultiHead}(Q, K, V) =& \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O \\ \text{head}_i =& \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{3–2}$$

其中Concat代表在通道维进行拼接，$h$ 代表注意力头的数量，$d_k = d_v = d_{model}/h$，$W_i^Q \in \mathbb{R}^{d_{model} \times d_k}$，$W_i^K \in \mathbb{R}^{d_{model} \times d_k}$，$W_i^V \in \mathbb{R}^{d_{model} \times d_v}$，$W_i^O \in \mathbb{R}^{h \cdot d_v \times d_{model}}$。

前馈神经网络由一个输入特征为 $d_{model} = 512$，输出特征为 $d_{ff} = 2048$ 的线性层，一个 ReLU 层，以及一个输入特征为 $d_{ff} = 2048$，输出特征为 $d_{model} = 512$ 的线性层堆叠而成。

在本文中，我们首先将前端提取到的特征传入一维卷积层，将降采样过后的特征加上一个 `[CLS]` 嵌入，一起送入 Transformer 编码器。`[CLS]` 是一个随机初始化的可学习的特殊嵌入，我们在每个视频片段的识别中都使用同一个 `[CLS]` 嵌入，其目的是整合输入的视频帧特征，这样我们可以使用 Transformer 编码器输出的 `[CLS]` 嵌入的上下文表征作为视频片段的整体表征。注意在输入到 Transformer 编码器之前，我们通过简单的绝对位置编码来添加帧之间的位置信息

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}) \end{aligned} \tag{3–3}$$

其中 $pos$ 是帧的位置，$i$ 是维度。绝对位置编码的每个维度对应于一个正弦波。波长形成一个从 $2\pi$ 到 $10000 \cdot 2\pi$ 的几何级数。

我们最初尝试使用 Stafylakis et al.[3]提出的 0-1 表示来将词边界整合到时序建模中，使用增广特征 $\mathbf{x}^+$ 作为后端的输入

$$\mathbf{x_t}^+ = [\mathbf{x_t}, b_t] \tag{3–4}$$

$\mathbf{x_t}$ 是第 $t$ 帧图像提取得到的特征，其中

$$b_t = \mathbb{1}_{t \in B} \tag{3–5}$$

$B$ 为所有目标词所在帧的集合。

经过实验，我们发现这种 0-1 指示函数在基于注意力机制的时序建模模型上效果欠佳。

从注意力机制本身着手，我们在 Transformer 编码器的自注意力（self attention）计算中使用了如图3–7所示的注意力遮罩 $M$

$$M = [M_{\texttt{[CLS]}}, M_1, \cdots, M_t, \ldots, M_T]^{T+1} \tag{3–6}$$

$M_t$ 为第 $t$ 帧对其他帧的注意力遮罩

| | <CLS> | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| <CLS> | -inf | -inf | -inf | 0 | 0 | -inf | -inf |
| 1 | -inf | 0 | -inf | -inf | -inf | -inf | -inf |
| 2 | -inf | -inf | 0 | -inf | -inf | -inf | -inf |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | -inf | -inf | -inf | -inf | -inf | 0 | -inf |
| 6 | -inf | -inf | -inf | -inf | -inf | -inf | 0 |

**图 3–7 提出的注意力遮罩方式。阴影部分表示注意力矩阵中被遮盖掉的部分，图示以 6 帧输入，第 3，4 帧为目标词所在帧为例**

$$M_t = \begin{cases} [-\inf, -\inf, -\inf, \cdots, -\inf]^{T+1}, & t\text{为}[\mathrm{CLS}] \\ [0, 0, 0, \cdots, 0]^{T+1}, & t \in B \\ [m_{[\mathrm{CLS}]}, m_0, m_1, \cdots, m_T]^{T+1}, & t \notin B \end{cases} \tag{3–7}$$

其中 $T = 29$ 为输入的视频片段帧数，对于 $t \notin B$ 的情况，$M_t$ 中第 $i$ 个元素 $m_i$ 为

$$m_i = \begin{cases} 0 & , i = t \\ -\inf & , i \neq t \end{cases} \tag{3–8}$$

这种注意力遮罩可以使在词边界外的帧只能看到自己，相当于其在 Transformer 编码器中不进行注意力机制计算，只进行前馈计算。目标词所在帧可以看到所有帧，从而进行时序建模，并且整合来自词边界外帧的关于特定视频片段的特征，如说话人特征，光照条件特征等，还可以整合词边界外帧的上下文信息。[CLS] 嵌入所在帧可以看到目标词所在帧，其可以整合这些帧的特征，输出能够代表整个视频片段的表征。

### 3.1.3 分类器

分类器是将解码器解码得到的整个视频片段表征进行分类的模块，我们使用了一个简单的一维卷积模块作为分类器，其包括一个输入通道数为 $d_{model} = 512$，输出通道数为 $d_{model} = 512$，步长为 1，卷积核大小为 1 的一维卷积层，批量归一化和 ReLU，以及最后一个输入通道数为 $d_{model} = 512$，输出通道数为 $N = 500$，步长为 1，卷积核大小为 1 的一维卷积层将特征投影到 $N = 500$ 个类别。分类器的输入为 [CLS] 嵌入经过解码器之后的输出，其能够代表经过时序建模模块整合后整个视频片段的特征。

### 3.1.4 解码器

除了分类器，我们还增加了一个解码器用于辅助的 CTC 损失计算，其结构类似于分类器中的一维卷积模块，但我们在其中将批归一化替换为层归一化[58]，层归一化更加适合对

于序列数据进行归一化处理，且其中第二个一维卷积层将特征投影到符号集大小 12。解码器的输入为

$$\mathbf{x} = \{\mathbf{x}_t\}, t \in B \tag{3–9}$$

其中 $\mathbf{x}_t$ 代表第 $t$ 个视频帧经过解码器的输出

### 3.1.5 损失函数

损失函数部分我们使用交叉熵损失和 CTC 损失的结合，交叉熵损失 $\mathcal{L}_{\text{CE}}$ 定义为

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^{N} y_{o,c} \log(p_{o,c}) \tag{3–10}$$

其中 $y_{o,c}$ 为指示类别 $c$ 是否为观测样本 $o$ 真实类别的二值化表示，$p_{o,c}$ 为观测样本 $o$ 属于类别 $c$ 的预测概率。CTC 损失 $\mathcal{L}_{\text{CTC}}$ 定义如下

$$\mathcal{L}_{\text{CTC}} = -\log p_{\text{CTC}}(\mathbf{y}|\mathbf{x}) \approx -\prod_{t=1}^{T} \log p(y_t|\mathbf{x}) \tag{3–11}$$

其中 $\mathbf{x} = [x_1, \cdots, x_T]$ 表示输入的视频帧，$\mathbf{y} = [y_1, \cdots, y_L]$ 表示视素类标签, 其中 $T$ 和 $L$ 分别表示输入帧序列和标签的长度，公式3–11中使用约等号是因为 CTC 损失假设每个输出预测之间的条件独立

我们整体使用的损失函数形式为

$$\mathcal{L} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{CE}} \tag{3–12}$$

其中 $\lambda$ 用于控制 CTC 损失和交叉损失的相对权重，我们的实现中使用 $\lambda = 0.6$。引入视素级 CTC 损失的好处是以视素为建模单元，给予解码器输出额外的监督信号，帮助其更好的学习目标词所在帧的特征，提升最终识别性能。使用视素是因为它是说话口型的最小视觉单位，是视频语音识别系统的最佳建模单元[59]。

## 3.2 实验

### 3.2.1 数据集

我们使用一个大规模的公开可用的词级唇读数据集 LRW[2]作为我们的测试平台，该数据集由 157 个小时视频组成，共计 489K 个来自 BBC 节目的视频片段，都有固定的 29 帧长度，以 25FPS 采样。每个视频片段都包含了 500 个词汇中的一个孤立的词，目标词出现在视频片段的正中间，周围有共同发音。所有的视频都是正面或接近正面的。

### 3.2.2 实验设置

我们使用视素级别的识别，符号集大小为 12，由 11 个视素以及用于 CTC 损失的特殊标记 `[blank]` 组成。由于数据集的转录只包含一个词，所以我们在词汇表中不包括空格、任何数字或标点符号。

我们使用 Park[60]来将 26 个英文字符先转化为 39 个 CMU 音素（phonemes），之后使用 Jeffers et al.[61]来将音素映射到 11 个视素，注意到每个英文单词都可以找到其音素序列，且音素到视素的映射是多对一的[62]。。具体的音素到视素映射表见表3–2。

| 视素 | 音素 |
|---|---|
| A | F V |
| B | ER OW R W UH UW |
| C | B P M |
| D | AW |
| E | DH TH |
| F | CH JH SH ZH |
| G | OY AO |
| H | S Z |
| I | AA AE AH AY EH EY IH IY Y |
| J | D L N T |
| K | G K NG HH |

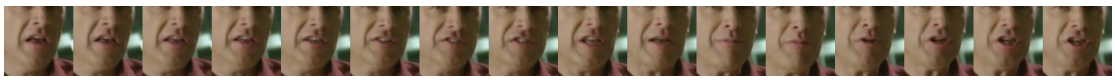
**表 3–2 CMU 音素到视素的映射**

我们的实现基于 Pytorch 库[63]，利用 Pytorch Lightning[64]实现，我们还整理了工具链用于复现本文的结果，具体参见第五章。我们使用四个 NVIDIA GeForce RTX 3090 GPU 进行训练，内存总量为 96GB，训练时长约为 2 天。网络使用 AdamW 优化器[65]进行训练，使用 $10^{-2}$ 大小的权重衰减（weight decay）来和概率为 0.2 的丢弃（dropout）进行正则化。初始学习率 $\eta$ 为 $3 \times 10^{-4}$。实验中模型的表现用准确率来表示。

a) 由 dlib 检测到的关键点。绿色的点表示 68 个脸部关键点，没有关键点的帧是 dlib 检测失败的

b) 线性插值后的脸部关键点

c) 用宽度为 12 的窗口对人脸进行平滑处理，并使用相似性变换将其与平均参考脸部关键点对齐



d) 使用 $120 \times 120$ 的边界框（bounding box）裁剪出唇部感兴趣区域

**图 3–8 用于说明生成唇部感兴趣区域过程的预处理实例**

**预处理：**我们对每段视频使用 dlib[66]检测并跟踪了 68 个脸部关键点。为了消除脸部旋

转和缩放造成的差异，我们使用 Martínez et al.[39]中的相似性变换将脸部与平均参考脸部关键点对齐。为了处理 dlib 无法检测到关键点的帧，我们使用了窗口宽度为 12 帧的关键点插值和平滑处理。最后使用一个以唇部区域为中心，大小为 $120 \times 120$ 的边界框来裁剪感兴趣区域。裁剪后的帧被转换为灰度，并根据训练集的总体平均值和方差进行归一化处理。预处理步骤样例参见图3–8

**数据增强：**在训练过程中按照 Ma et al.[16]中的设置，我们对视频片段使用 $112 \times 112$ 的随机裁剪和概率为 0.5 的水平翻转。

**标签平滑：**标签平滑是一种通过添加标签噪声来提高模型泛化能力的一种常见做法，它具有惩罚低熵输出（即过度自信的预测）分布的效果。Pereyra et al.[67]展示了这种策略在几个常见的基准任务上的有效性，包括图像分类、语言建模、机器翻译和语音识别。

交叉熵损失的标签平滑方式为

$$y'_{o,c} = \begin{cases} \epsilon/N & ,c\text{不为}o\text{真实标签} \\ 1 - \dfrac{N-1}{N}\epsilon & ,c\text{为}o\text{真实标签} \end{cases} \tag{3–13}$$

其中 $N = 500$ 为类别数，$\epsilon$ 是一个小常数，在我们的实现中交叉熵损失的 $\epsilon = 0.1$。

Kim et al.[68]提出了用于 CTC 损失的标签平滑策略，通过在 CTC 损失的目标函数中加入了一个正则化项来实现，这个正则化项由网络的预测分布 $P$ 和标签间的均匀分布 $U$ 之间的 KL 散度组成

$$\mathcal{L} = \epsilon \sum_{t=1}^{T} D_{KL}(P_t||U) + (1-\epsilon)\mathcal{L}_{\text{CTC}} \tag{3–14}$$

其中对于 CTC 损失 $\epsilon = 0.01$

**余弦学习率调度器：**余弦学习率调度器使得模型可以从大的学习率开始训练，经过最大周期数 $T$ 降低到一个最小值。在训练过程中降低学习率有助于帮助模型的收敛。

$$\eta_t = \eta_{min} + \frac{1}{2}(\eta - \eta_{min})(1 + \cos(\frac{t}{T}\pi)) \tag{3–15}$$

其中 $\eta_t$ 为第 $t$ 个时期（epoch）的学习率，$\eta = 3 \times 10^{-4}$ 为起始学习率，$\eta_{min}$ 为经过 $T$ 个时期后学习率下降到的最小值

**学习率预热** 学习率预热是一种减少早期训练样本的主要影响（primacy effect）的方法。如果没有它，我们可能需要运行一些额外的时期来使得模型收敛，因为模型会解除对那些早期迷信的训练。我们在余弦学习率调度器前增加一个 $T_{warmup} = 10$ 个周期的线性学习率预热阶段，学习率在预热期内线性增加

$$\eta_t = \frac{t}{T_{warmup}}\eta \tag{3–16}$$

### 3.2.3 实验结果

我们在表3–5中介绍实验的结果，报告了提出的模型的准确率。我们提出的词级视频语音识别模型实现了 89.1% 的准确率，比目前最先进的模型 Kim et al.[9]提高了 0.6%。这是一个相当大的改进，该任务从 Feng et al.[4]于 2020 年达到了 88.4% 的准确率之后一直没有较大提升。我们的模型相较目前最好的基于注意力机制的 Wiriyathammabhum[7]准确率绝对提升了 4.7%，验证了在词级视频语音识别这个短序列分类任务上基于注意力机制的模型可以很

好地利用词边界信息达到突出的表现。另外注意到 Kim et al.[9]需要使用标注的对齐的音视频进行训练，而我们的自监督预训练阶段只需要使用无标注的单模图像进行训练即可，数据要求更简单。

### 3.2.4 消融实验

| 方法 | 准确率（%） |
|---|---|
| 基线 | 87.4 |
| + ResNet-50 前端 | 87.8 |
| + MoCo v2 前端 | 88.8 |
| + 字符级 CTC 损失 | 89.0 |
| + 视素级 CTC 损失 | 89.1 |

**表 3–3 词级视频语音识别模型消融实验**

模型消融实验的结果展示在表3–3中。基线模型使用 ResNet-18 而不是 MoCo v2 作为前端，不使用 CTC 辅助损失。我们首先用 ResNet-50 前端替换 ResNet-18 前端，观察到了 0.4% 的准确率提升。然后我们用预训练的 MoCo v2 权重初始化 ResNet-50 前端，提升了 1.0% 的准确率，这验证了使用自监督预训练视觉模型可以有效改进视频语音识别任务。之后加上辅助的字符级 CTC 损失来帮助模型中时序建模模块学习到更好的表征，我们观察到了 0.2% 的提升。最后，我们通过引入视素作为建模单元，因为视素是视频语音识别更为自然的建模单元，我们又获得了 89.1% 的提升。

### 3.2.5 音频模态语音识别

| 方法 | 准确率（%） |
|---|---|
| MFCC + 双向 GRU[42] | 97.7 |
| 一维 ResNet-18 + 双向 GRU[42] | 97.7 |
| 一维 ResNet-18 + 双向 LSTM + 词边界[3] | 98.6 |
| 一维 ResNet-18 + Transformer + 词边界 + 视素 CTC 损失（我们） | **98.7** |

**表 3–4 词级语音识别模型在 LRW 数据集上的分类准确率**

我们还测试了 Transformer 后端以及辅助 CTC 损失在音频模态上的性能，结果如表3–4所示，我们的模型达到了 98.7% 的准确率，是目前 LRW 数据集上表现最好的语音识别模型，这在音频模态验证了我们在后端使用的注意力遮罩以及辅助 CTC 损失的效果。音频模型使用采样率为 16kHz 的原始音频作为输入，通过一个简单的一维 ResNet-18 前端得到 25FPS 的音频帧特征。之后类似于视觉模态，使用 Transformer 和注意力遮罩进行解码，使用分类器进行分类，并使用解码器辅助学习。注意在音频模态我们使用音素作为建模单元，音素通过相同方法生成，共有 39 类。

## 3.3 本章小结

本章介绍了我们提出的词级视频语音识别模型。本章首先介绍了模型的整体架构以及模型的特征维度，之后分模块介绍了前端、后端、分类器和解码器四个部分以及使用的损失函数。在前端部分我们介绍了前端模型，还介绍了使用 MoCo v2 预训练模型进行特征提取存在的问题，以及提出的解决方案。在后端部分，我们详细介绍了使用的 Transformer 编码器的架构和计算方法，还介绍了使用的注意力遮罩的生成方式。分类器和解码器部分中介绍了使用的一维卷积模型。损失函数部分，我们介绍了使用的交叉熵损失和 CTC 损失，以及他们的计算方法。在实验部分，我们介绍了实验中详细的设置，以及我们的实验结果和消融实验。本章主要贡献在于提出了注意力遮罩来整合词边界的方法，以及使用 CTC 损失辅助训练词级视频语音识别任务的想法，最终准确率超过目前 LRW 数据集上最优模型 0.6%，验证了我们提出方法的有效性。

| 方法 | 前端 | 后端 | 准确率（%） |
|---|---|---|---|
| LRW[2] | VGG-M | - | 61.1 |
| ResNet18 + 时序卷积网络[38] | 三维卷积 + ResNet18 | 时序卷积网络 | 74.6 |
| ResNet50 + 时序卷积网络 | 三维卷积 + ResNet50 | 时序卷积网络 | 76.7 |
| MoCo v2 + 时序卷积网络 | 三维卷积 + MoCo v2 | 时序卷积网络 | 79.0 |
| ResNet34 + 双向 LSTM[38] | 三维卷积 + ResNet34 | 双向 LSTM | 83.0 |
| ResNet34 + 双向 GRU[42] | 三维卷积 + ResNet34 | 双向 GRU | 83.4 |
| SpotFast + Transformer[7] | SpotFast 网络 | Lateral Transformers | 84.4 |
| MoCo v2 + Transformer | 三维卷积 + MoCo v2 | Transformer | 85.0 |
| 多尺度时序卷积网络[39] | 三维卷积 + ResNet18 | 多尺度时序卷积网络 | 85.3 |
| ResNet-18 + 双向 LSTM + 词边界[3] | 三维卷积 + SE-ResNet-18 | 双向 LSTM | 88.1 |
| 稠密时序卷积网络[5] | 三维卷积 + ResNet18 | 稠密时序卷积网络 | 88.4 |
| SE-ResNet-18 + 双向 GRU + 词边界[4] | 三维卷积 + SE-ResNet-18 | 双向 GRU | 88.4 |
| 多尺度时序卷积网络 + 知识蒸馏（集成模型）[45] | 三维卷积 + ResNet18 | 多尺度时序卷积网络 | 88.5 |
| 三维卷积 + ResNet-18 + 多尺度时序卷积网络 + 多头视觉-音频记忆（自监督模型）[9] | 三维卷积 + ResNet-18 | 多尺度时序卷积网络 + 多头视觉-音频记忆 | 88.5 |
| MoCo v2 + Transformer + 词边界 + 视素 CTC 损失（我们） | 三维卷积 + MoCo v2 | Transformer | **89.1** |

**表 3–5 词级视频语音识别模型在 LRW 数据集上的分类准确率**

# 第四章　句级音视频语音识别

## 4.1　模型

本章节将对我们提出的句级音视频语音识别模型进行介绍。模型主要分为音频和视觉两个模态的前端和后端，融合模块和解码器。音频和视觉前端分别通过预训练的 wav2vec 2.0 和 MoCo v2 进行特征提取，单模的后端由 Transformer 分别进行单模时序建模，然后在融合模块将两个模态提取得的特征进行融合并进行多模时序建模，再经过 seq2seq 和 CTC 两个解码器完成解码。

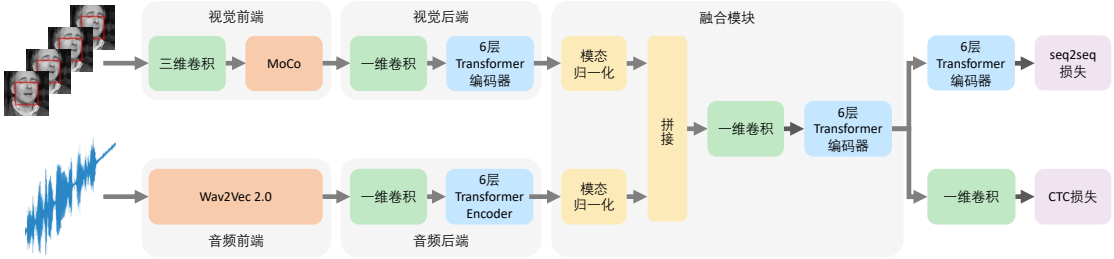



**图 4–1　句级音视频语音识别模型架构**

图4–1中展示了句级音视频语音识别模型的整体架构。模型的输入是灰度视频序列，视觉部分是经过裁剪得到的感兴趣区域，帧率为 25FPS，音频部分使用原始音频（raw audio）作为输入，采样率为 16kHz。

### 4.1.1　前端

| 部分 | 模块 | 音频输入 $T_s \times 1$ |
|---|---|---|
| 前端 | wav2vec 2.0 | $T_f \times 1024$ |
| 后端 | 一维卷积 | $\frac{T_f}{2} \times d_{model}$ |
| | Transformer 编码器 | $\frac{T_f}{2} \times d_{model}$ |

**表 4–1　音频流的特征维度**
**{时间维大小 × 通道数} 代表特征维度，$T_s$ 和 $T_f$ 分别表示采样的音频输入和音频帧的数量，$d_{model} = 512$ 代表模型维度。**

**音频前端：**音频前端是相当简单的。我们使用在 Libri-Light[17]上预训练过的 wav2vec 2.0[27]。按照它被应用于自动语音识别任务的一般范式，我们在音频前端中使用其中的一维卷积和 Transformer 编码器。音频前端将 16kHz 的原始音频作为输入，wav2vec 2.0 中的一维卷积模块窗口宽度为 25ms，步长为 20ms，通过这个一维卷积模块，16kHz 的原始音频输入被降频为 49Hz 的音频帧。详细的音频流特征尺寸见表4–1。

| 部分 | 模块 | 视频输入<br>$T_f \times 112^2 \times 1$ |
|---|---|---|
| 前端 | 三维卷积 | $T_f \times 28^2 \times 64$ |
| | MoCo v2 | $T_f \times 2048$ |
| 后端 | 一维卷积 | $T_f \times d_{model}$ |
| | Transformer 编码器 | $T_f \times d_{model}$ |

**表 4–2 视频流的特征维度**
**{时间维大小 × (空间维大小$^2$) × 通道数} 代表特征维度，$T_f$ 表示视觉帧的数量，**
**$d_{model} = 512$ 代表模型维度。**

**视觉前端：**本章使用的视觉前端与3.1.1节中介绍的 MoCo v2 前端相同。如表4–2所示，视频输入为 112×112 大小的灰度图像序列，采样率为 25FPS。需要注意的是在此任务中，由于视觉模态直接学习句级语音识别难度较高，我们首先在第三章中介绍的词级视频语音识别任务上对视觉前端进行预训练，以更好地表示唇部动作。预训练使用的模型为图3–5中所示的使用调整后的 MoCo v2 作为前端的词级视频语音识别模型，训练收敛后我们只将前端迁移到此任务中作为视觉前端。

### 4.1.2 后端

由于视频帧为 25FPS，而原始音频经过前端处理后得到的音频帧为 49Hz 左右。注意到音视频两个模态的频率有 2 倍左右的差异。由于 wav2vec 2.0 中一维卷积的感受野（receptive field）较大，音视频两个模态的频率比例并不是正好 2 倍。我们通过适当地对音频序列在前后进行复制填充，或截断尾部的音频帧来将音频模态频率调整为 50Hz，保证视觉帧和音频帧之间恰好 1:2 的比例。在后端，我们在时间维度上使用一维卷积层使两个模态具有相同的频率，结合 Transformer 编码器提供单一模态的时序建模。

**音频后端：**在音频后端，输入为频率 50Hz，特征维度 1024 的音频帧。我们通过一维卷积来将频率降低为 25Hz 同时特征维度降为 $d_{model} = 512$，使得音频帧能与视频帧在时间维度上一一对应。一维卷积包括一个输入通道数为 1024，输出通道数为 $d_{model} = 512$，步长为 2，卷积核大小为 2 的一维卷积层，以及层归一化和 ReLU。我们使用一个 6 层 Transform 编码器，其前馈维度 $d_{ff} = 2048$。对于位置编码，我们使用公式3–3中定义的绝对位置编码。表4–1清晰展示了音频流的输入输出尺寸。

**视觉后端：**视觉后端的输入为频率 25Hz，特征维度 2048 的视频帧。在视觉后端，我们对应使用一个一维卷积来保持这个频率同时将特征维度降为 $d_{model} = 512$，其包括一个输入通道数为 2048，输出通道数为 $d_{model} = 512$，步长为 1，卷积核大小为 1 的一维卷积层，以及层归一化和 ReLU。Transformer 编码器和绝对位置编码的设置与音频后端相同，但注意视觉后端中的 Transformer 编码器另外使用一套单独的参数，而非与音频后端共享参数。详细的视频流输入输出尺寸见表4–2。

### 4.1.3 融合模块

来自音频和视觉模态的特征在融合模块被融合在一起，我们在通道维度将其进行拼接，形成频率为 25Hz，特征维度为 1024 的音视频帧。在将每个模态的特征维度连接起来之前，我们对每个模态分别使用不带仿射变换的层归一化[58]，这可以避免其中某个一个模态的方差较大，使得模型忽略了来自另一模态的特征。我们使用与视觉后端中类似的一维卷积来将音视频帧的特征维度降低至 $d_{model}$ = 512，随后的 6 层 Transformer 编码器也与音频和视觉两个模态后端中的 Transformer 编码器设置相同，其可以帮助模型进行多模时序建模。

### 4.1.4 解码器

我们按照 Petridis et al.[10]的设置，同时使用 seq2seq 和 CTC 两个解码器基于融合模块的相同输出同时进行训练。

**seq2seq 解码器：**第一个是 seq2seq 解码器，使用了一个 6 层 Transformer 解码器（decoder）和一维卷积。

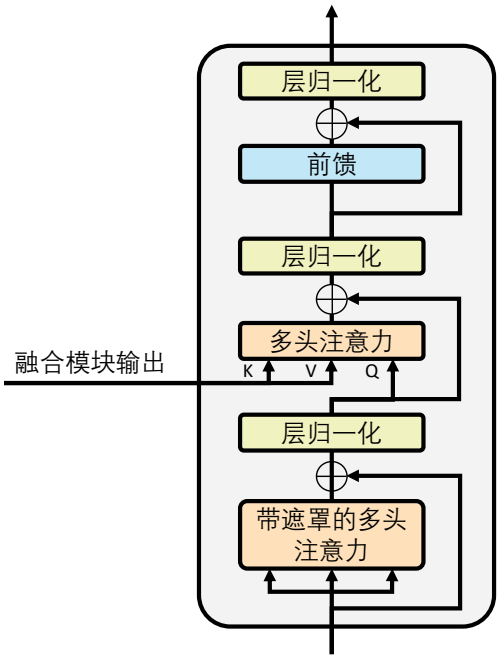



**图 4–2 Transformer 解码器层**

Transformer 解码器在 Vaswani et al.[46]中提出。每个 Transformer 解码器层的结构如图4–2所示，其中包含两个多头注意层，注意力机制的计算同3.1.2节中介绍的 Transformer 编码器中的注意力机制。

第一个为带遮罩的多头注意力层，其输入同时作为注意力的查询，键和值。其注意力遮罩如图4–3所示，其下三角部分为 0，其余部分为-inf，这种设计使得输入的每个嵌入只能看到自己以及之前的嵌入，而之后的信息不会被泄露。这允许在训练 Transformer 解码器时进行并行的教师强制（teacher forcing）训练，通过使用完整的真实标签作为输入，先经过词嵌入模块将其投射到嵌入向量，之后经过公式3–3中定义的绝对位置编码加入位置信息，由于存在注意力遮罩的设计，每个词嵌入的输出可以并行得到。在推理（inference）过程中使用

| | <SOS> | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| <SOS> | 0 | -inf | -inf | -inf | -inf |
| 1 | 0 | 0 | -inf | -inf | -inf |
| 2 | 0 | 0 | 0 | -inf | -inf |
| 3 | 0 | 0 | 0 | 0 | -inf |
| 4 | 0 | 0 | 0 | 0 | 0 |

**图 4–3 Transformer 解码器注意力遮罩**

自回归（autoregression），输入可以通过连接之前预测得到的前缀和最后一个时间步骤中的假设字符来获得，推理过程将无法并行计算得到。

第二个多头注意力层使用第一个注意力层得到的特征作为查询，使用来自融合模块的输出的音视频特征作为键和值。这个注意力层使得由来自已解码出文本的特征，对音视频特征进行查询，整合其中信息得到可以预测下一个时间步输出的特征。

最后的一维卷积由两个步长和卷积核大小均为 1 的一维卷积层组成，两层之间有层归一化和 ReLU。第一个卷积层输入输出通道数都为 $d_{model} = 512$，第二个卷积层输入通道数为 $d_{model} = 512$，输出通道数为符号集大小 40。因为我们使用字符级别的预测，大小为 40 的符号集由字母表中的 26 个字符、10 个数字、撇号和以下特殊标记组成：用于隔开单词的 `[space]`、用于 CTC 损失的 `[blank]` 和表示句子开始和结束的 `[EOS/SOS]`。由于数据集中不包含其他标点符号，我们不把它们列入词汇表。

**CTC 解码器：**第二个是 CTC 解码器，是一个一维卷积模块，其设置与 seq2seq 解码器中相同。

### 4.1.5 损失函数

在这项工作中，我们使用了混合 CTC/注意力损失[69]。设 $\mathbf{x} = [x_1, \ldots, x_T]$ 为融合模块的输入音视频帧序列，$\mathbf{y} = [y_1, \ldots, y_L]$ 为标签，其中 $T$ 和 $L$ 分别表示输入和标签的长度。CTC 损失假定每个输出预测之间的条件独立性，其形式为

$$\mathcal{L}_{\text{CTC}} = -\log p_{\text{CTC}}(\mathbf{y}|\mathbf{x}) \approx -\prod_{t=1}^{T} \log p(y_t|\mathbf{x}) \tag{4–1}$$

另一方面，seq2seq 解码器通过在链式规则的基础上直接估计后验来摆脱这一假设，其形式为

$$\mathcal{L}_{\text{CE}} = -\log p_{\text{CE}}(\mathbf{y}|\mathbf{x}) = -\prod_{l=1}^{L} \log p(y_l|y_{<l}, \mathbf{x}) \tag{4–2}$$

总体损失的计算方法如下

$$\mathcal{L} = \lambda \mathcal{L}_{\text{CTC}} + (1-\lambda)\mathcal{L}_{\text{CE}} \tag{4–3}$$

其中 $\lambda$ 控制 CTC 损失和交叉熵损失在混合 CTC/注意力机制中的相对权重，在我们的实现中 $\lambda = 0.2$。这个权重不仅用于将两个损失整合为一个训练损失，而且在解码过程中也被用作融合 CTC 和 seq2seq 两个解码器的预测，我们将在下面的4.1.6节中重新讨论这个问题。使用混合 CTC/注意力损失的好处是注意力机制的引入可以帮助消除 CTC 损失的条件独立假设。

### 4.1.6 解码

解码采用 Watanabe et al.[69]中提出的 CTC/注意力联合单程解码（joint CTC/attention one-pass decoding）与波束搜索（beam search）算法。我们应用浅层融合（shallow fusion）来融合 CTC 和 seq2seq 的预测：

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}\in\hat{\mathcal{Y}}} \{\alpha \log p_{\mathrm{CTC}}(\mathbf{y}|\mathbf{x}) + (1-\alpha) \log p_{\mathrm{CE}}(\mathbf{y}|\mathbf{x})\} \tag{4–4}$$

其中 $\hat{\mathcal{Y}}$ 表示目标符号的预测集，$\alpha$ 是在验证集上调整的相对权重，在我们的实现中 $\alpha = 0.1$。使用混合 CTC/注意力解码的好处是，通过 CTC 概率作为正则项，减少纯注意力解码带来的不规则对齐输出的数量。

**Algorithm 4–1** 混合 CTC/注意力单程解码算法，改编自 Watanabe et al.[69]。
记号: $X$ 表示输入; $L_{max}$ 为搜索空间中假设的最大长度，我们将其设置为 $T$; $C$ 为解码得到的符号序列; `[b]` 表示 `[blank]`.

**输入:** $X, L_{max}$
**输出:** $C$

$\Omega_0 = \{\texttt{[SOS]}\}$
$\hat{\Omega} = \emptyset$
$\gamma_0^{(b)}(\texttt{[SOS]}) = 1$
**for** $t = 1, \cdots, T$ **do**
    $\gamma_t^{(n)}(\texttt{[SOS]}) = 0$
    $\gamma_t^{(b)}(\texttt{[SOS]}) = \prod_{\tau=1}^{t} \gamma_{\tau-1}^{(b)}(\texttt{[SOS]}) \cdot p(z_\tau = \texttt{[b]}|X)$
**end for**
**for** $l = 1 \cdots L_{max}$ **do**
    $\Omega_l = \emptyset$
    **while** $\Omega_{l-1} \neq \emptyset$ **do**
        $g = \mathrm{HEAD}(\Omega_{l-1})$
        DEQUEUE($\Omega_{l-1}$)
        **for** each $c \in \mathcal{U}$ **do**
            $h = g \cdot c$
            **if** $c = \texttt{[EOS]}$ **then**
                $\log p_{\mathrm{ctc}}(h|X) = \log\{\gamma_T^{(n)}(g) + \gamma_T^{(b)}(g)\}$
            **else**
                **if** $g = \texttt{[SOS]}$ **then**
                    $\gamma_1^{(n)}(h) = p(z_1 = c|X)$
                **else**
                    $\gamma_1^{(n)}(h) = 0$

**end if**
$\gamma_1^{(b)}(h) = 0$
$\Psi = \gamma_1^{(n)}(h)$
**for** $t = 2 \cdots T$ **do**
**if** $last(g) = c$ **then**
$\Phi = \gamma_{t-1}^{(b)}(g)$
**else**
$\Phi = \gamma_{t-1}^{(b)}(g) + \gamma_{t-1}^{(n)}(g)$
**end if**
$\gamma_t^{(n)}(h) = (\gamma_{t-1}^{(n)}(h) + \Phi) p(z_t = c|X)$
$\gamma_t^{(b)}(h) = (\gamma_{t-1}^{(b)}(h) + \gamma_{t-1}^{(n)}(h)) p(z_t = \texttt{[b]}|X)$
$\Psi = \Psi + \Phi \cdot p(z_t = c|X)$
**end for**
$\log p_{\text{ctc}}(h|X) = \log(\Psi)$
**end if**
$\log p(h|X) = \alpha \log p_{\text{ctc}}(h|X) + (1 - \alpha) \log p_{\text{att}}(h|X)$
**if** $c =$ `[EOS]` **then**
ENQUEUE$(\hat{\Omega}, h)$
**else**
ENQUEUE$(\Omega_l, h)$
**end if**
**end for**
**end while**
$\Omega_l = \text{TOPK}(\Omega_l, W)$
**end for**
**return** $\arg\max_{C \in \hat{\Omega}} \log p(C|X)$

算法4–1描述了混合 CTC/注意力解码算法。CTC 前缀概率被定义为所有以 $h$ 为前缀的符号序列的累积概率：

$$p_{\text{ctc}}(h|X) = \sum_{v \in (\mathcal{U})^+} p_{\text{ctc}}(h \cdot v|X) \tag{4–5}$$

其中 $v$ 表示所有可能的非空符号序列。CTC 概率可以通过前向假设概率 $\gamma_t^{(n)}$ 和 $\gamma_t^{(b)}$ 来计算，其中上标 $(n)$ 和 $(b)$ 分别代表所有 CTC 路径以非 `[blank]` 或 `[blank]` 符号结束。

解码算法也集成了宽度为 $W$ 的波束搜索，其中超参数 $\alpha$ 控制给予 CTC 和注意力解码的相对权重。$\mathcal{U}$ 是除 `[blank]` 外的符号集，在我们的实现中，`[SOS]` 和 `[EOS]` 使用同一符号代替。

### 4.1.7 训练流水线

最终的句级音视频语音识别模型是通过一个训练流水线实现的。

对于音频模态，首先通过自监督学习对音频前端进行预训练，这是由 wav2vec 2.0 完成的。然后，音频前端在纯音频（AO: audio only）设置下和随机初始化的音频后端以及解码器一起进行训练。

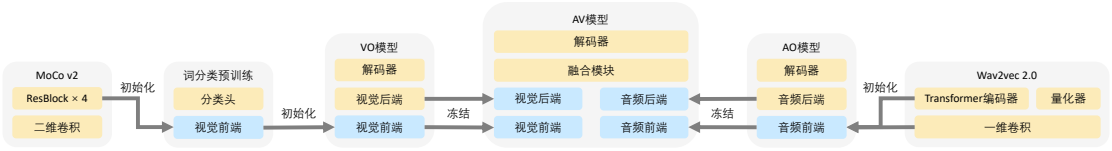



**图 4–4 训练流水线**
**黄色模块代表随机初始化的新参数，而蓝色模块代表从上一个训练阶段继承的参数**

对于视觉模态，视觉前端首先通过自监督学习进行预训练，然后通过词级视频语音识别任务进行训练。之后，视觉前端由纯视觉（VO: visual only）模型继承，其中使用随机初始化的视觉后端和解码器。

最终的音视频（AV: audio visual）模型可以在纯音频和纯视觉模型收敛后进行训练。由于计算资源的限制，我们预先计算了音频和视觉后端输出，并在最后阶段只学习融合模块和解码器部分的参数。我们的训练管道的详细可视化描述见图4–4。

## 4.2 实验

### 4.2.1 数据集

我们使用大规模的公开句级音视频语音识别数据集 LRS2[70]（Lip Reading Sentences 2）作为我们的主要测试平台。在训练过程中，我们还使用3.2.1节中介绍的 LRW 数据集，通过加入额外的的词级视频语音识别任务对视觉前端进行预训练。

LRS2 由 224 个小时的对齐的音频和视频组成，总共有 144K 个来自 BBC 视频的片段，这些片段的长度为句子级别。训练数据包含超过 200 万个单词和共计 4 万种词汇。该数据集非常具有挑战性，因为说话人的头部姿势、口音和照明条件都非常多样。

### 4.2.2 实验设置

同3.2.2节中介绍的实现相同，我们基于 Pytorch 库[63]，利用 Pytorch Lightning[64]实现了本模型。相关工具链也可用于复现本部分结果，具体参见第五章。我们在四个 NVIDIA A100 GPU（内存总量为 160GB）上进行了约一周的训练。网络使用 Adam 优化器[71]进行训练，初始学习率为 $10^{-4}$。我们也使用了3.2.2中介绍的标签平滑策略，seq2seq 损失和 CTC 损失的标签平滑 $\epsilon$ 都设置为 0.01。我们使用了与3.2.2节中相同的学习率预热策略，但在预热期结束后通过 ReduceOnplateau 学习率调度器进行学习调节。由于 LRS2 预训练集中样本长度较长，我们随机抽取占整个样本 1/3 单词数的连续片段，以便与训练集中的片段长度相匹配。若抽取得到的片段长度超过 160 帧，我们继续减少单词数直到其长度符合要求。

**预处理与数据增强：**本章使用与3.2.2节中相同的视觉预处理和数据增强方法。对于音频模态，在训练纯音频和音视频模型时，我们通过在原始音频时域上添加加性多路重合噪声（babble noise）来增强噪声鲁棒性，同时在音视频模型中，添加噪声还可以增大模型学习音频模态信息的难度，有助于模型学习到来自视觉模态的信息。我们的实现中添加噪声的信噪比为 5dB，概率为 $p_n = 0.25$。多路重合噪音是通过混合 LRS2 中随机选取的 20 个不同音频样本合成的。

**评估方法：**对于本章中的实验，我们通过错词率（WER: word error rate）来衡量模型的

表现，其定义为

$$\mathrm{WER} = (S + D + I)/N \tag{4–6}$$

其中的 $S$、$D$ 和 $I$ 分别表示从真实标签到预测结果的编辑距离（edit distance）中替换、删除和插入的数量，$N$ 是标签中的词数。

在评估过程中，添加到音频波形中的多路重合噪声是用与训练相同的方式产生的，但我们设置了一个不同的种子以避免模型拟合到特定的合成噪声。解码使用4.1.6中介绍的混合 CTC/注意力单程解码算法，波束宽度为 $W = 5$[①]。我们在实验中不使用外部语言模型（language model）。

**ReduceOnplateau 调度器：**本章使用 ReduceOnplateau 调度器来调节学习率。在验证集的错词率经过一定时期没有继续降低时，调度器会将学习率降低为原先一半，模型表现通常会得到提升并开始继续学习。

### 4.2.3 实验结果

我们在表4–4中介绍了所有实验的结果，报告了纯视觉、纯音频和音视频三种设置下的错词率。请注意，这里列出的许多模型在训练流水线的不同阶段中也使用了额外的训练数据，如 MV-LRS[13]、LRS3[12]、LibriSpeech[72]和 LRW[2]。

| 模块 | 我们的模型 | TM-CTC | E2E Conformer |
|---|---|---|---|
| 音频前端 | 315.0M | - | 3.9M |
| 视觉前端 | 23.5M | 11.2M（冻结） | 11.2M |
| 音频后端 | 20.2M | 20.2M | 31.8M |
| 视觉后端 | 20.2M | 20.2M | 31.8M |
| 融合模块 | 19.7M | 19.7M | 0.8M |
| 解码器 | 26.2M | 20.5K | 9.5M |

**表 4–3 我们的模型、TM-CTC[1]和 E2E Conformer[16]模型的参数量大小比较**

我们将我们的模型、TM-CTC 模型[1]和当前最先进的模型 E2E Conformer[16]的参数量大小列于表4–3中。我们的模型后端和融合模块配置遵循 TM-CTC 模型，seq2seq 解码器中的超参数设置与后端相同。最重要的区别是，我们利用了预先训练好的前端，这带来了更大的模型规模。

**音视频设置：**在主要的音视频设置下，我们使用 LRS2 中的预训练和训练集作为最终训练阶段的训练集。我们提出的句级音视频语音识别模型在没有任何外部语言模型的帮助下达到了 2.6% 的误码率，比目前最先进的 Ma et al.[16]提高了 1.1%。这是一个相当大的改进，相对改进约为 30%。

**纯音频设置：**用于训练纯音频模型的训练数据包括来自 LRS2 的 224 小时标注数据，以及来自 Libri-Light[17]的 60K 小时未标注数据，这些数据通过继承 wav2vec 2.0 中的参数被间接使用。我们的模型在纯音频设置下实现了 2.7% 的错词率，这相比目前最优模型 Ma et al.[16]的错词率降低了 1.2%，相对改进了 31%。

① 通过在 LRS2 的验证集上实验确定

| 方法 | 错词率（%） |
|---|---|
| 纯音频 | |
| TM-CTC*[1] | 10.1 |
| TM-seq2seq*[1] | 9.7 |
| CTC/attention*[10] | 8.2 |
| LF-MMI TDNN*[73] | 6.7 |
| E2E Conformer**[16] | 3.9 |
| 我们的模型 | **2.7** |
| 纯视觉 | |
| LIBS[74] | 65.3 |
| TM-CTC*[1] | 54.7 |
| Conv-seq2seq[11] | 51.7 |
| TM-seq2seq*[1] | 50.0 |
| KD-TM[75] | 49.2 |
| LF-MMI TDNN*[73] | 48.9 |
| E2E Conformer*[16] | 42.4 |
| E2E Conformer**[16] | **37.9** |
| 我们的模型 | 43.2 |
| 音视频 | |
| TM-DCM[48] | 8.6 |
| TM-seq2seq*[1] | 8.5 |
| TM-CTC*[1] | 8.2 |
| LF-MMI TDNN*[73] | 5.9 |
| E2E Conformer**[16] | 3.7 |
| 我们的模型 | **2.6** |

**表 4–4 在 LRS2 上测试的纯音频、纯视觉和音视频模型的错词率结果**
**带 * 的模型表示结果使用了外部语言模型，这表明在其评估过程中比我们的模型有优势**
**带 ** 的模型表示它使用了一个更强大的基于 Transformer 的语言模型**

**纯视觉设置：**纯视觉模型使用 LRS2 中的预训练和训练集进行训练。在词级视频语音识别预训练中我们额外使用了 LRW 数据集，同时还通过继承 MoCo v2 的参数间接使用了 ImageNet 的 1.28M 张未标注的图像。与目前最优模型 E2E Conformer[16]相比，主要区别是其在解码过程中使用了一个大型的 Transformer 语言模型，与其消融实验中使用的基于循环神经网络的语言模型相比提升了 4.5% 的错词率。我们的纯视觉模型和带有基于循环神经网络语言模型的 E2E Conformer 模型之间的差距是 0.8%，这是一个合理的范围。此外，我们使用 6 层 Transformer 编码器进行时序建模，而 E2E Conformer 使用了为语音识别任务定制的 12 层 Conformer[52]编码器，在时序建模部分比我们更加具有优势。如果我们只考虑与不使用外部语言模型的工作来进行更公平的比较，目前最优模型是 Ren et al.[75]，其错词率为

49.2%，比我们的模型落后 6.0%。

**解码样例：**表4–5中展示了纯音频模型不能预测而音视频模型可以正确预测的句子示例。视觉模态的引入可以帮助模型克服多样的错误类型。

| |
|---|
| *AO:* WHATEVER YOU ASK<br>*AV:* WHATEVER YOU ARE |
| *AO:* TRAVEL THREE MILES URBER WEST AND YOU DO GET MORE FOR YOUR MONEY HERE<br>*AV:* TRAVEL THREE MILES FURTHER WEST AND YOU DO GET MORE FOR YOUR MONEY HERE |
| *AO:* IT COULD BE YOUR PASSPORT FOR A SMALL FORTUNE<br>*AV:* IT COULD BE YOUR PASSPORT TO A SMALL FORTUNE |
| *AO:* WHAT TO THINK FOR THEMSELVES<br>*AV:* NOT TO THINK FOR THEMSELVES |
| *AO:* NOT THE SUBJECT MATTERING<br>*AV:* NOT FOR SUBJECT MATTER |
| *AO:* I WOULDN'T SAY I'M THE STAR<br>*AV:* I WOULDN'T SAY I'M A STAR |
| *AO:* CRISPAS PUDDING THAT NOBODY REALLY LIKES<br>*AV:* CHRISTMAS PUDDING THAT NOBODY REALLY LIKES |
| *AO:* BUT AT THE SAME TIME<br>*AV:* AT THE SAME TIME |
| *AO:* BEING ~~ON~~ MY OWN<br>*AV:* BEING MY OWN |
| *AO:* ~~SO~~ AT ONE POINT<br>*AV:* AT ONE POINT |

**表 4–5 AO（纯音频）和 AV（音视频）设置解码样例**
**下划线表示替换和插入错误，删除线表示删除错误**

### 4.2.4 消融实验

在本节中，我们在纯音频和纯视觉设置中测试单独的组件对模型性能的影响。

**纯音频：**LRS2 上的纯音频模型的消融结果显示在表4–6中。从 Afouras et al.[1]开始，我们首先使用在 LibriSpeech 上预训练的 wav2vec 2.0 前端替换 STFT 音频特征，结果提升了 11.1%。然后，我们将其替换为在更大的无标注单模音频数据集 Libri-Light 上预训练的 wav2vec 2.0 前端，观察到 0.6% 的进一步提升。我们进一步在训练阶段使用混合 CTC/注意力进行优化和解码，又有了 0.9% 的提升。

**纯视觉：**LRS2 上的纯视觉模型的结果显示在表4–7。从 Afouras et al.[1]开始，我们首先通过使用混合 CTC/注意力引入端到端训练（前端仍然预先通过 LRW 进行预训练，但不使用课程学习（curriculum learning）），结果提高了 16.0%。然后，我们使用预训练的 MoCo v2

| 方法 | 错词率（%） |
|---|---|
| 基线[1] | 15.3 |
| + wav2vec 2.0 （LibriSpeech）前端 | 4.2 |
| + wav2vec 2.0 （Libri-Light）前端 | 3.6 |
| + 混合 CTC/注意力 | **2.7** |

**表 4–6 LRS2 上的纯音频模型性能的消融实验**

| 方法 | 错词率（%） |
|---|---|
| 基线[1] | 65.0 |
| + 混合 CTC/注意力 | 49.0 |
| + MoCo v2 前端 | **43.2** |

**表 4–7 LRS2 上的纯视觉模型性能的消融实验**

权重初始化前端，带来了 5.8% 的进一步改进。

### 4.2.5 噪声环境下的鲁棒性

为了评估模型对音频噪声的耐受性，我们测试了我们的模型在不同信噪比水平的多路重合噪声下的表现。

| 设置 | 模型 | 0dB | 5dB | 无噪声 |
|---|---|---|---|---|
| 纯音频 | Afouras et al.[1] | 58.0% | - | 10.5% |
| | 我们的模型 | **32.5**% | 6.8% | **2.7**% |
| 音视频 | Afouras et al.[1] | 33.5% | - | 9.4% |
| | 我们的模型 | **24.5**% | 6.3% | **2.6**% |

**表 4–8 不同信噪比水平下的错词率，噪声是合成的多路重合噪声**

如表4–8所示，当信噪比水平为 0dB 时，我们的纯音频和音视频模型的错词率分别达到了 32.5% 和 24.5%，相比 Afouras et al.[1]中的结果降低了 25.5% 和 9%①。当信噪比水平上升到 5dB 时，我们的纯音频和视听模型达到了 6.8% 和 6.3% 的错词率。

除了在多路重合噪声环境下取得了比基线模型更显著的改善，我们进一步研究了模型在人类噪声环境下的表现。人类噪音是非常具有挑战性的，因为噪音本身包含一些单词，而模型不容易区分应该识别哪个音频信号。我们通过从 LRS2 数据集中的不同音频样本中随机裁剪出许多 1 秒钟的信号来合成人类噪音。如图所示，我们进行了不同噪声条件下的人类噪声的实验，模型是用多路重合噪声增强过的音频进行训练的。当信噪比水平下降至 0db 以下时，错词率大大增加。这是因为在低信噪比水平下，模型可能无法分辨出两个重叠的口语单词。并且我们发现在每个信噪比水平下的其整体表现都比多路重合噪声差，这说明在带有特定信息的噪声下进行语音识别比无序的多路重合噪声下更难。

① Ma et al.[16]也提供了噪声环境下的错词率，然而由于论文中缺乏必要的细节，我们无法生成相同的噪声并与其进行比较。

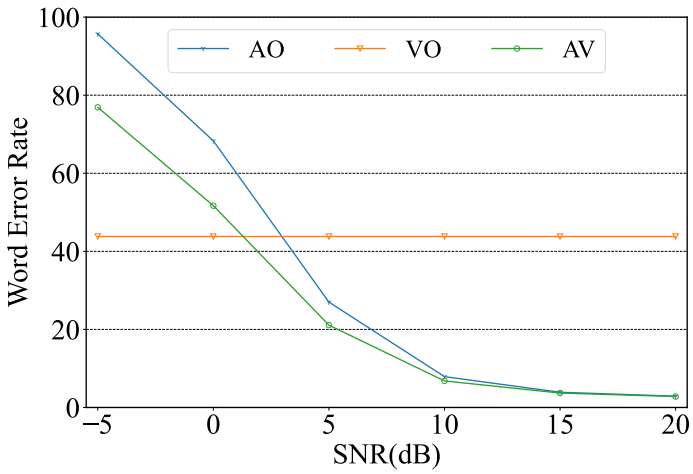



**图 4–5 不同的信噪比水平下的错词率，噪声是 LRS2 中采样的人类语音**
**AO：纯音频模型, VO：纯视觉模型, AV：音视频模型**

| 设置 | 训练数据（小时数） | 错词率（%） |
|---|---|---|
| 纯音频 | LRS2 （224） | 2.7 |
| | LRS2 训练集（28） | 3.4 (+0.7) |
| 纯视觉 | LRS2 （224） | 43.2 |
| | LRS2 训练集（28） | 68.9 (+25.7) |

**表 4–9 使用不同规模训练数据的纯音频和纯视觉模型性能**

### 4.2.6 低资源下的语音识别

使用自监督预训练模型的一个重要好处是只需要少量的标注数据来训练模型。

为了进一步研究模型在低资源环境下的表现，我们使用 LRS2 的 28 小时训练集来训练纯音频和纯视觉模型，结果报告在表4–9中。用 28 小时数据训练的纯音频模型的错词率为 3.4%，相比用 224 小时数据训练的模型稍差，但仍然超过目前最优的 Ma et al.[16]0.3%。这一结果表明，对于纯音频模型来说，在大规模单模数据集上预训练的自监督模型可以显著降低对数据的要求。而用 28 小时数据训练的纯视觉模型与用 224 小时数据训练的模型有很大差距，原因可能是纯视觉模型更难训练，需要更多的数据量。

### 4.2.7 讨论

我们发现音频模态下使用自监督模型比视觉模态下使用自监督模型有更大的改进。我们认为其原因可以列举如下：

1. 音频模态的训练数据规模明显大于视觉模态，用于预训练 wav2vec 2.0 的 Libri-Light 数据集由 60K 小时的音频信号组成，相反，用于预训练 MoCo v2 的 ImageNet 数据集只有 1.28M 张图片，大致相当于 14 小时 25FPS 下的无声视频。
2. MoCo v2 模型对图像进行预训练，特征可以很好地表示单帧图像内容，但这无法建模视觉帧之间的时间相关性。相比之下，wav2vec 2.0 模型在原始音频上进行了预训

练，对音频帧之间的时间相关性进行了建模，因此具有更好的时序建模能力。

## 4.3 本章小结

本章介绍了我们提出的句级音视频语音识别模型，首先介绍了模型的整体架构，之后分模块介绍了前端、后端、融合模块和解码器以及我们使用的损失函数和解码算法以及训练流程。在前后端部分，我们介绍了两个模态的特征维度，以及 Transformer 解码器的架构。损失函数和解码部分介绍了使用的混合 CTC/注意力机制，我们还在训练流程部分详细说明了纯音频、纯视觉和音视频三个模型的获得方法。实验部分我们介绍了详细的实验细节，以及三种设置下的最终结果，提供了解码样例。我们还展示了消融实验结果以证明使用单模自监督预训练的作用。对比了纯音频和音视频两种设置在不同噪声环境下的识别准确率。还探索了我们提出的模型在人声噪音和低资源两种条件下的效果。本章的主要贡献在于提出使用单模自监督预训练来提升多模音视频语音识别任务的性能，在 LRS2 数据集上取得了 2.6% 的错词率，相对改进了 30%。

# 第五章　音视频语音识别工具链

目前多模音视频语音识别领域相关的开源工作较少，第一个使用了深度神经网络的现代音视频语音识别模型[1]以及目前 LRS2 数据集上最先进的 Ma et al.[16]都没有开源代码。词级视频语音识别领域，最先进的 Kim et al.[9]仅开源了测试代码。目前本领域的开源工作主要为 Martínez et al.[39]和 Ma et al.[45]，以及 Stafylakis et al.[3]的词级视频语音识别开源代码仓库，缺少可供参考的成熟音视频语音识别开源方案，缺少相关工具链。本章我们拓展了第三章和第四章两部分工作的实现，初步搭建了音视频语音识别工具链。

## 5.1　设计总览

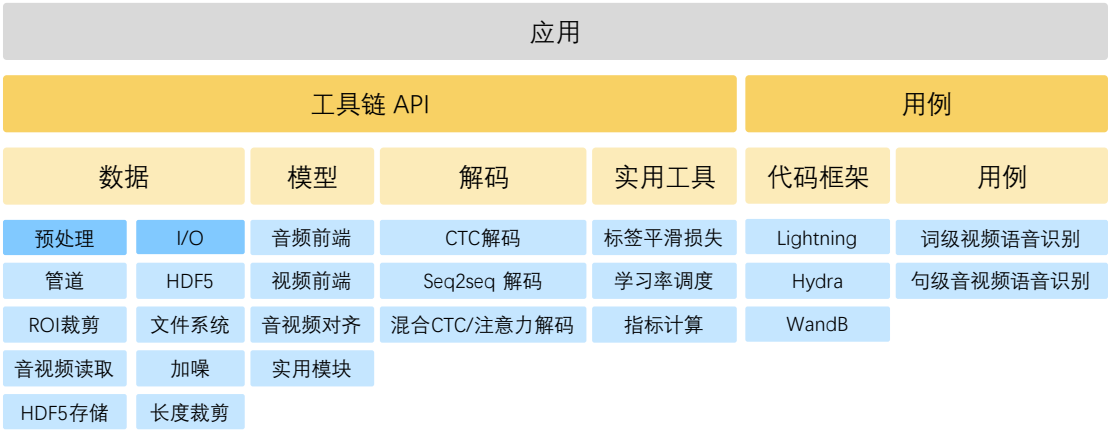



**图 5–1　音视频语音识别工具链设计**

如图5–1所示，工具链主要由工具链 API 和用例两部分组成。工具链 API 部分提供了构建音视频语音识别任务中常用的预定义模块。用例部分提供了训练代码框架，以及使用工具链 API 和代码框架实现的第三章和第四章中提出的模型。

## 5.2　工具链 API

工具链 API 部分由数据、模型、解码、实用工具四部分组成。主要提供了用于快速搭建多模音视频语音识别所需的预定义组件

### 5.2.1　数据

数据部分由预处理和 I/O 两个子部分组成。分别有预定义的数据管道，使用时通过简单组合就可以实现复杂的功能。

**预处理：**在预处理部分，我们定义了预处理管道。如图5–2所示，通过简单的加入音频和视频的预处理设置定义预处理数据管道，传入 mp4 文件的路径就可以对其进行预处理。

预定义的感兴趣区域裁剪模块支持灵活的自定义功能，无论是选择使用3.2.2节中描述的 dlib 裁剪算法，还是使用默认的感兴趣区域裁剪坐标都可以轻松的通过更改参数实现。

```python
from avsr.preprocess import preprocessors

preprocessors = preprocessors.Compose(
    [
        preprocessors.ProcessAudio(
            channel=1,
            rate=16000
        ),
        preprocessors.ProcessVisual(
            dlib_detect=True,
            mean_face_landmarks=None,
            mean_face_files_list=files_list,
            mean_face_generate_n_process=8,
            landmark_detector=landmark_detector,
            frame_size=(224, 224),
            crop_size=(120, 120),
            default_crop=None),
    ]
)
preprocessors('/path/to/*.mp4')
```

**图 5–2 使用工具链对 mp4 文件进行预处理**

我们还支持多线程预处理，并提供了相应的基于 Python multiprocessing 库的预处理脚本，方便使用者在其基础上自定义修改。

**数据 I/O：**数据 I/O 部分在本工具链中主要体现为训练时的数据载入部分。我们定义了数据读取和处理管道，用于在 Pytorch 数据集中的 __getitem__ 函数处调用。

```python
import h5py
import numpy as np

from avsr.data import datapipes

h5 = h5py.File('/path/to/h5_file', 'r')
noise = np.randn(10000, )
datapipe = datapipes.Compose(
    [
        datapipes.ReadAudioFromh5(h5=h5, subset='train', normalize=True),
        datapipes.AddNoise(noise=noise, noise_prob=0.25, noise_SNR=5),
        datapipes.ReadVisualFromh5(h5=h5, subset='train', shape=(120, 120)),
        datapipes.CropVideoClip(max_length=160)
    ]
)
sample = datapipe({'index': index})
```

**图 5–3 使用数据管道在 __getitem__ 函数中从 HDF5 读取数据并进行处理**

如图5–3所示，我们通过简单地构建数据管道，可以实现音视频读取、加噪、4.2.2节中描述的视频长度裁剪等功能。数据管道还内置了更加方便的从 mp4 或 numpy 文件中读取视频，从波形文件中读取音频的功能，但因为效率不如 HDF5 高，所以不推荐用户使用。

此外我们还拓展了 torchvision[76]中的 transforms 组件，使其能够支持对视频进行批量的归一化、随机裁剪、随机翻转、中心裁剪等操作。

### 5.2.2 模型

模型部分工具链主要定义了3.1.1节中描述的 MoCo v2 视觉前端，4.1.1节中描述的 wav2vec 2.0 音频前端。还提供了4.1.2节中介绍的音视频对齐方案。此外我们还提供了一维 ResNet-18，三维 ResNet-18 的实现，以及生成遮罩矩阵等实用模块。

### 5.2.3 解码

解码部分我们提供了三种解码算法的实现，分别是4.1.6节中介绍的混合 CTC/注意力解码算法，贪心 CTC 解码算法和 seq2seq 自回归解码算法。

### 5.2.4 实用工具

使用工具部分我们提供了很多预定义的模块, 用于拓展原生 Pytorch 的实现。包括3.2.2节中介绍的标签平滑交叉熵损失和 CTC 损失，3.2.2节中使用的准确率和4.2.2节中使用的错词率和错字率计算以及学习率预热 ReduceLROnPlateau 联合调度器。

## 5.3 用例

### 5.3.1 代码框架

我们使用 Pytorch Lightning[64]、Hydra[77]、WandB[78]搭建了通用的训练代码框架。

**Pytorch Lightning：** PyTorch Lightning[64]是面向专业人工智能研究人员和机器学习工程师的深度学习框架，其封装了训练代码，将算法部分暴露出来，在使用上非常灵活。并且其模型和数据类都继承自原生 Pytorch，可以轻松完成从 Pytorch 的迁移。我们选择使用 Lightning 主要是因为其灵活性很大，不会因为抽象程度过高而干扰研究人员的使用。Lightning 功能强大，提供了非常便捷的分布式数据并行（distributed data parallel）多卡训练方式, 可以自动记录训练过程中的指标, 保存检查点, 支持从检查点恢复训练。

**Hydra：** Hydra[77]是一款配置工具，允许用户从 YAML 文件和命令行中动态组合配置。还可以从命令行中覆盖 YAML 的配置。这使得我们在使用 YAML 文件管理配置的同时，不必再编写复杂的 argparse 指令来完成参数的增加以及覆盖。

**WandB：** WandB[78]是一款机器学习实验追踪工具，可以帮助用户更快地建立更好的模型。其可以记录每次训练使用的超参数，训练过程中的指标变化，并自动上传检查点至云端，这可以帮助用户复现实验结果。

### 5.3.2 用例

用例部分我们使用工具链实现了第三章中描述的词级视频语音识别模型和第四章中描述的句级音视频语音识别模型，还提供了使用工具链进行 LRW 和 LRS2 两个数据集预处理的脚本。我们还提供了根据5.3.1节中介绍的代码框架搭建的示例训练、测试和推理代码。用户可以通过本节部分轻松复现我们的实验结果，或者以我们的代码为基线，快速构建出自定义模型。

## 5.4 本章小结

本章介绍了我们构建的音视频语音识别工具链，旨在帮助研究人员更方便地复现我们的工作和进行多模音视频语音识别相关的研究。我们介绍了工具链的设计总览，分模块介绍了工具链的 API 设计和用例实现。

# 全文总结

## 本文工作总结

本文探索了基于预训练模型的多模语音识别系统，着重解决在多模音视频语音识别任务中应用自监督预训练模型的两大难点。其一，对于如何在视频语音识别任务中应用预训练模型的问题，我们通过将通用视觉自监督预训练模型中的二维卷积更换为随机初始化的三维卷积，赋予预训练特征提取器捕获视觉帧之间时间相关性的能力，同时降低了无监督预训练数据域间差异带来的影响。其二，对于如何将这些预训练模型整合到多模场景中的问题，我们将预训练过的音频和视觉前端的部分参数整合到一个多模音视频语音识别框架中，通过一整套训练流水线得到最终的音视频模型。

我们使用该领域最为流行的两个任务：词级视频语音识别和句级音视频语音识别任务来对我们的方法进行测试。在词级任务中，我们提出了在基于注意力机制的模型中使用注意力遮罩来整合词边界信息的方法，在目标词所在帧通过注意力整合词边界外帧的环境和上下文信息，通过特殊的 `[CLS]` 嵌入整合目标词所在帧的信息。使得基于注意力机制的模型可以有效利用词边界信息提升词级视频语音识别任务性能。我们还使用视素级 CTC 损失给予解码器输出额外的视素级监督信号，帮助优化目标词所在帧的特征。最终我们提出的模型在大规模公开数据集 LRW 上取得了 89.1% 的新的最优分类准确率性能，较目前最优模型绝对提升了 0.6%。在句级任务中，我们提出的框架通过混合 CTC/注意力训练及解码将对齐的音视频输入识别为文字，框架中从单模自监督学习中继承的两个前端合作良好，多模框架可以通过微调产生有竞争力的结果。即使没有外部语言模型，我们提出的模型也大幅提高了大规模公开数据集 LRS2 上音视频语音识别任务的表现，达到了 2.6% 的新的最优错词率，相对目前最优模型提高了 30%。

## 未来工作

本节基于第三章和第四章中提出的模型。总结了未来可以继续探索和尝试的方向。

### 基于 Transformer 的特征提取模块

在本文设计的两部分工作中，我们均使用基于 MoCo v2 的视觉特征提取器，其架构为 ResNet[79]。最近基于注意力机制的 Transformer 视觉特征提取器如 ViT[80]在很多视觉任务如分类、检测等上取得了突出效果。使用基于注意力机制的视觉特征提取器使得模型能够选择输入图像中对视频语音识别最为重要的信息。

### 单模自监督预训练

4.2.7节中介绍了本文使用的 MoCo v2 预训练视觉前端不如 wav2vec 2.0 预训练音频前端的原因：视觉模态的训练数据较少，MoCo v2 不对时间相关性进行建模。未来可以通过使用更大规模的数据集进行自监督预训练进行改善，使用人脸数据集预期将会缩小与视频语音识别的域间差异，带来额外提升。关于时间相关性建模的缺失，我们希望引入视频级预训练模型来提高模型建模帧间时间相关性的能力。

多模自监督预训练

第四章中提出了使用两个单模预训练模型来提升多模音视频语音识别任务性能的方法。最近出现了很多多模自监督预训练模型，如使用互联网中图像文本对进行训练的视觉语言表示模型 CLIP[81]，使用图像的自然语言描述（image caption）数据集进行训练的视觉语言表示模型 ViLT[82]等。其中 Kim et al.[82]提出了使用多模态 Transformer 进行多模自监督预训练的方法，我们认为未来可以使用成对的音频视觉对进行多模自监督预训练，得到更好的单多模表示。

# 参考文献

[1] AFOURAS T, CHUNG J, SENIOR A, et al. Deep Audio-visual Speech Recognition[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2018: 1-1. DOI: 10.1109/TPAMI .2018.2889052.

[2] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]//Asian conference on computer vision. [S.l. : s.n.], 2016: 87-103.

[3] STAFYLAKIS T, KHAN M H, TZIMIROPOULOS G. Pushing the boundaries of audiovisual word recognition using residual networks and LSTMs[J]. Computer Vision and Image Understanding, 2018, 176: 22-32.

[4] FENG D, YANG S, SHAN S, et al. Learn an effective lip reading model without pains[J/OL]. ArXiv preprint, 2020, abs/2011.07557. https://arxiv.org/abs/2011.07557.

[5] MA P, WANG Y, SHEN J, et al. Lip-reading with densely connected temporal convolutional networks[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. [S.l. : s.n.], 2021: 2857-2866.

[6] YANG S, ZHANG Y, FENG D, et al. LRW-1000: A naturally-distributed large-scale benchmark for lip reading in the wild[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). [S.l. : s.n.], 2019: 1-8.

[7] WIRIYATHAMMABHUM P. SpotFast Networks with Memory Augmented Lateral Transformers for Lipreading[C]//International Conference on Neural Information Processing. [S.l. : s.n.], 2020: 554-561.

[8] CHEN X, FAN H, GIRSHICK R, et al. Improved baselines with momentum contrastive learning[J/OL]. ArXiv preprint, 2020, abs/2003.04297. https://arxiv.org/abs/2003.04297.

[9] KIM M, YEO J H, RO Y M. Distinguishing Homophenes Using Multi-Head Visual-Audio Memory for Lip Reading[C]//36th AAAI Conference on Artificial Intelligence (AAAI 22). [S.l. : s.n.], 2022.

[10] PETRIDIS S, STAFYLAKIS T, MA P, et al. Audio-visual speech recognition with a hybrid ctc/attention architecture[C]//2018 IEEE Spoken Language Technology Workshop (SLT). [S.l. : s.n.], 2018: 513-520.

[11] ZHANG X, CHENG F, WANG S. Spatio-Temporal Fusion Based Convolutional Sequence Learning for Lip Reading[C/OL]//2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019. [S.l.]: IEEE, 2019: 713-722. https://doi.org/10.1109/ICCV.2019.00080. DOI: 10.1109/ICCV.2019.00080.

[12] AFOURAS T, CHUNG J S, ZISSERMAN A. LRS3-TED: a large-scale dataset for visual speech recognition[J/OL]. ArXiv preprint, 2018, abs/1809.00496. https://arxiv.org/abs/1809.00496.

[13] CHUNG J S, ZISSERMAN A. Lip Reading in Profile[C/OL]//British Machine Vision Conference 2017, BMVC 2017, London, UK, September 4-7, 2017. [S.l.]: BMVA Press, 2017. https://www.dropbox.com/s/20a1cgndopwk7e7/0706.pdf?dl=1.

[14] AFOURAS T, CHUNG J S, ZISSERMAN A. ASR is All You Need: Cross-Modal Distillation for Lip Reading[C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. [S.l.]: IEEE, 2020: 2143-2147. https://doi.org/10.1109/ICASSP40776.2020.9054253. DOI: 10.1109/ICASSP40776.2020.9054253.

[15] CHUNG J S, NAGRANI A, ZISSERMAN A. VoxCeleb2: Deep Speaker Recognition[C/OL]//YEGNANARAYANA B. Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. [S.l.]: ISCA, 2018: 1086-1090. https://doi.org/10.21437/Interspeech.2018-1929. DOI: 10.21437/Interspeech.2018-1929.

[16] MA P, PETRIDIS S, PANTIC M. End-To-End Audio-Visual Speech Recognition with Conformers[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l. : s.n.], 2021: 7613-7617.

[17] KAHN J, RIVIÈRE M, ZHENG W, et al. Libri-Light: A Benchmark for ASR with Limited or No Supervision[C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. [S.l.]: IEEE, 2020: 7669-7673. https://doi.org/10.1109/ICASSP40776.2020.9052942. DOI: 10.1109/ICASSP40776.2020.9052942.

[18] DENG J, DONG W, SOCHER R, et al. ImageNet: A large-scale hierarchical image database[C/OL]//2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA. [S.l.]: IEEE Computer Society, 2009: 248-255. https://doi.org/10.1109/CVPR.2009.5206848. DOI: 10.1109/CVPR.2009.5206848.

[19] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft coco: Common objects in context[C]//European conference on computer vision. [S.l. : s.n.], 2014: 740-755.

[20] BROWN T B, MANN B, RYDER N, et al. Language Models are Few-Shot Learners[C/OL]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. [S.l. : s.n.], 2020. https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[21] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C/OL]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, Minnesota: Association for Computational Linguistics, 2019: 4171-4186. https://aclanthology.org/N19-1423. DOI: 10.18653/v1/N19-1423.

[22] BAEVSKI A, ZHOU Y, MOHAMED A, et al. Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations[C/OL]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. [S.l. : s.n.], 2020. https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

[23] HE K, FAN H, WU Y, et al. Momentum Contrast for Unsupervised Visual Representation Learning. CoRR abs/1911.05722 (2019)[J/OL]. ArXiv preprint, 2019, abs/1911.05722. https://arxiv.org/abs/1911.05722.

[24] CHEN T, KORNBLITH S, NOROUZI M, et al. A Simple Framework for Contrastive Learning of Visual Representations[C/OL]//Proceedings of Machine Learning Research: Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event: vol. 119. [S.l.]: PMLR, 2020: 1597-1607. http://proceedings.mlr.press/v119/chen20j.html.

[25] GRILL J, STRUB F, ALTCHÉ F, et al. Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning[C/OL]//LAROCHELLE H, RANZATO M, HADSELL R, et al. Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual. [S.l. : s.n.], 2020. https://proceedings.neurips.cc/paper/2020/hash/f3ada80d5c4ee70142b17b8192b2958e-Abstract.html.

[26] SHUKLA A, VOUGIOUKAS K, MA P, et al. Visually Guided Self Supervised Learning of Speech Representations[C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. [S.l.]: IEEE, 2020: 6299-6303. https://doi.org/10.1109/ICASSP40776.2020.9053415. DOI: 10.1109/ICASSP40776.2020.9053415.

[27] SCHNEIDER S, BAEVSKI A, COLLOBERT R, et al. Wav2vec: Unsupervised Pre-Training for Speech Recognition[C/OL]//KUBIN G, KACIC Z. Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019. [S.l.]: ISCA, 2019: 3465-3469. https://doi.org/10.21437/Interspeech.2019-1873. DOI: 10.21437/Interspeech.2019-1873.

[28] BAEVSKI A, SCHNEIDER S, AULI M. Vq-wav2vec: Self-Supervised Learning of Discrete Speech Representations[C/OL]//8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. [S.l.]: OpenReview.net, 2020. https://openreview.net/forum?id=rylwJxrYDS.

[29] CHEN S, WANG C, CHEN Z, et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing[J/OL]. ArXiv preprint, 2021, abs/2110.13900. https://arxiv.org/abs/2110.13900.

[30] HSU W N, BOLTE B, TSAI Y H H, et al. Hubert: Self-supervised speech representation learning by masked prediction of hidden units[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3451-3460.

[31] CHEN X, HE K. Exploring Simple Siamese Representation Learning[J/OL]. ArXiv preprint, 2020, abs/2011.10566. https://arxiv.org/abs/2011.10566.

[32] LÜSCHER C, BECK E, IRIE K, et al. RWTH ASR Systems for LibriSpeech: Hybrid vs Attention - w/o Data Augmentation[C]//INTERSPEECH. [S.l. : s.n.], 2019.

[33] AMODEI D, ANANTHANARAYANAN S, ANUBHAI R, et al. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin[C/OL]//BALCAN M, WEINBERGER K Q. JMLR Workshop and Conference Proceedings: Proceedings of the 33nd International

Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016: vol. 48. [S.l.]: JMLR.org, 2016: 173-182. http://proceedings.mlr.press/v48/amodei16.html.

[34] PALAZ D, MAGIMAI-DOSS M, COLLOBERT R. Convolutional Neural Networks-based continuous speech recognition using raw speech signal[C/OL]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. [S.l.]: IEEE, 2015: 4295-4299. https://doi.org/10.1109/ICASSP.2015.7178781. DOI: 10.1109/ICASSP.2015.7178781.

[35] ZEGHIDOUR N, XU Q, LIPTCHINSKY V, et al. Fully Convolutional Speech Recognition[J]. ArXiv, 2018, abs/1812.06864.

[36] GRAVES A, FERNÁNDEZ S, GOMEZ F J, et al. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks[C/OL]//COHEN W W, MOORE A W. ACM International Conference Proceeding Series: Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006: vol. 148. [S.l.]: ACM, 2006: 369-376. https://doi.org/10.1145/1143844.1143891. DOI: 10.1145/1143844.1143891.

[37] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/Attention Architecture for End-to-End Speech Recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253. DOI: 10.1109/JSTSP.2017.2763455.

[38] STAFYLAKIS T, TZIMIROPOULOS G. Combining Residual Networks with LSTMs for Lipreading[C/OL]//LACERDA F. Interspeech 2017, 18th Annual Conference of the International Speech Communication Association, Stockholm, Sweden, August 20-24, 2017. [S.l.]: ISCA, 2017: 3652-3656. http://www.isca-speech.org/archive/Interspeech%5C_2017/abstracts/0085.html.

[39] MARTÍNEZ B, MA P, PETRIDIS S, et al. Lipreading Using Temporal Convolutional Networks[C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. [S.l.]: IEEE, 2020: 6319-6323. https://doi.org/10.1109/ICASSP40776.2020.9053841. DOI: 10.1109/ICASSP40776.2020.9053841.

[40] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[41] CHO K, van MERRIËNBOER B, GULCEHRE C, et al. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation[C/OL]//Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1724-1734. https://aclanthology.org/D14-1179. DOI: 10.3115/v1/D14-1179.

[42] PETRIDIS S, STAFYLAKIS T, MA P, et al. End-to-End Audiovisual Speech Recognition[C/OL]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2018, Calgary, AB, Canada, April 15-20, 2018. [S.l.]: IEEE, 2018: 6548-6552. https://doi.org/10.1109/ICASSP.2018.8461326. DOI: 10.1109/ICASSP.2018.8461326.

[43] ZHAO X, YANG S, SHAN S, et al. Mutual information maximization for effective lip reading[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). [S.l. : s.n.], 2020: 420-427.

[44] ZHANG Y, YANG S, XIAO J, et al. Can we read speech beyond the lips? rethinking roi selection for deep visual speech recognition[C]//2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). [S.l. : s.n.], 2020: 356-363.

[45] MA P, MARTINEZ B, PETRIDIS S, et al. Towards practical lipreading with distilled and efficient models[C]//ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S.l. : s.n.], 2021: 7608-7612.

[46] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C/OL]// GUYON I, von LUXBURG U, BENGIO S, et al. Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. [S.l. : s.n.], 2017: 5998-6008. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[47] DUPONT S, LUETTIN J. Audio-visual speech modeling for continuous speech recognition[J]. IEEE Transactions on Multimedia, 2000, 2(3): 141-151. DOI: 10.1109/6046.865479.

[48] LEE Y H, JANG D W, KIM J B, et al. Audio–visual speech recognition based on dual cross-modality attentions with the transformer model[J]. Applied Sciences, 2020, 10(20): 7263.

[49] LI W, WANG S, LEI M, et al. Improving Audio-visual Speech Recognition Performance with Cross-modal Student-teacher Training[C/OL]//IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019. [S.l.]: IEEE, 2019: 6560-6564. https://doi.org/10.1109/ICASSP.2019.8682868. DOI: 10.1109/ICASSP.2019.8682868.

[50] PARASKEVOPOULOS G, PARTHASARATHY S, KHARE A, et al. Multiresolution and multimodal speech recognition with transformers[J/OL]. ArXiv preprint, 2020, abs/2004.14840. https://arxiv.org/abs/2004.14840.

[51] TAO F, BUSSO C. End-to-end audiovisual speech recognition system with multitask learning[J]. IEEE Transactions on Multimedia, 2020, 23: 1-11.

[52] GULATI A, QIN J, CHIU C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[C/OL]//MENG H, XU B, ZHENG T F. Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020. [S.l.]: ISCA, 2020: 5036-5040. https://doi.org/10.21437/Interspeech.2020-3015. DOI: 10.21437/Interspeech.2020-3015.

[53] HADSELL R, CHOPRA S, LECUN Y. Dimensionality reduction by learning an invariant mapping[C]//2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06): vol. 2. [S.l. : s.n.], 2006: 1735-1742.

[54] GIDARIS S, SINGH P, KOMODAKIS N. Unsupervised Representation Learning by Predicting Image Rotations[C/OL]//6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. [S.l.]: OpenReview.net, 2018. https://openreview.net/forum?id=S1v4N2l0-.

[55] ZHANG R, ISOLA P, EFROS A A. Colorful image colorization[C]//European conference on computer vision. [S.l. : s.n.], 2016: 649-666.

[56] DOERSCH C, GUPTA A, EFROS A A. Unsupervised Visual Representation Learning by Context Prediction[C/OL]//2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. [S.l.]: IEEE Computer Society, 2015: 1422-1430. https://doi.org/10.1109/ICCV.2015.167. DOI: 10.1109/ICCV.2015.167.

[57] IOFFE S, SZEGEDY C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[C/OL]//BACH F R, BLEI D M. JMLR Workshop and Conference Proceedings: Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015: vol. 37. [S.l.]: JMLR.org, 2015: 448-456. http://proceedings.mlr.press/v37/ioffe15.html.

[58] BA J L, KIROS J R, HINTON G E. Layer normalization[J/OL]. ArXiv preprint, 2016, abs/1607.06450. https://arxiv.org/abs/1607.06450.

[59] FERNANDEZ-LOPEZ A, SUKNO F M. Optimizing Phoneme-to-Viseme Mapping for Continuous Lip-Reading in Spanish[C]//International Joint Conference on Computer Vision, Imaging and Computer Graphics. [S.l. : s.n.], 2017: 305-328.

[60] PARK J, Kyubyong & Kim. G2pE[Z]. https://github.com/Kyubyong/g2p. 2019.

[61] JEFFERS J, BARLEY M. Speechreading (lipreading)[M/OL]. [S.l.]: Thomas, 1971. https://books.google.com.sg/books?id=-UNlGOqM5eUC.

[62] ELLIOTT E A. Phonological Functions of Facial Movements: Evidence from deaf users of German Sign Language[J]. Thesis, 2013.

[63] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C/OL]//WALLACH H M, LAROCHELLE H, BEYGELZIMER A, et al. Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada. [S.l. : s.n.], 2019: 8024-8035. https://proceedings.neurips.cc/paper/2019/hash/bdbca288fee7f92f2bfa9f7012727740-Abstract.html.

[64] FALCON E A, WA. PyTorch Lightning[J]. GitHub. Note: https://github.com/PyTorchLightning/pytorch-lightning, 2019, 3.

[65] LOSHCHILOV I, HUTTER F. Decoupled Weight Decay Regularization[C/OL]//7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. [S.l.]: OpenReview.net, 2019. https://openreview.net/forum?id=Bkg6RiCqY7.

[66] KING D E. Dlib-ml: A machine learning toolkit[J]. The Journal of Machine Learning Research, 2009, 10: 1755-1758.

[67] PEREYRA G, TUCKER G, CHOROWSKI J, et al. Regularizing neural networks by penalizing confident output distributions[J/OL]. ArXiv preprint, 2017, abs/1701.06548. https://arxiv.org/abs/1701.06548.

[68] KIM S, SELTZER M L, LI J, et al. Improved Training for Online End-to-end Speech Recognition Systems[C/OL]//YEGNANARAYANA B. Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018. [S.l.]: ISCA, 2018: 2913-2917. https://doi.org/10.21437/Interspeech.2018-2517. DOI: 10.21437/Interspeech.2018-2517.

[69] WATANABE S, HORI T, KIM S, et al. Hybrid CTC/attention architecture for end-to-end speech recognition[J]. IEEE Journal of Selected Topics in Signal Processing, 2017, 11(8): 1240-1253.

[70] CHUNG J S, SENIOR A W, VINYALS O, et al. Lip Reading Sentences in the Wild[C/OL]//2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. [S.l.]: IEEE Computer Society, 2017: 3444-3453. https://doi.org/10.1109/CVPR.2017.367. DOI: 10.1109/CVPR.2017.367.

[71] KINGMA D P, BA J. Adam: A Method for Stochastic Optimization[C/OL]//BENGIO Y, LECUN Y. 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. [S.l. : s.n.], 2015. http://arxiv.org/abs/1412.6980.

[72] PANAYOTOV V, CHEN G, POVEY D, et al. Librispeech: An ASR corpus based on public domain audio books[C/OL]//2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015. [S.l.]: IEEE, 2015: 5206-5210. https://doi.org/10.1109/ICASSP.2015.7178964. DOI: 10.1109/ICASSP.2015.7178964.

[73] YU J, ZHANG S, WU J, et al. Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset[C/OL]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020. [S.l.]: IEEE, 2020: 6984-6988. https://doi.org/10.1109/ICASSP40776.2020.9054127. DOI: 10.1109/ICASSP40776.2020.9054127.

[74] ZHAO Y, XU R, WANG X, et al. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers[C/OL]//The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. [S.l.]: AAAI Press, 2020: 6917-6924. https://aaai.org/ojs/index.php/AAAI/article/view/6174.

[75] REN S, DU Y, LV J, et al. Learning From the Master: Distilling Cross-Modal Advanced Knowledge for Lip Reading[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l. : s.n.], 2021: 13325-13333.

[76] MARCEL S, RODRIGUEZ Y. Torchvision the machine-vision package of torch[C]//Proceedings of the 18th ACM international conference on Multimedia. [S.l. : s.n.], 2010: 1485-1488.

[77] YADAN O. Hydra - A framework for elegantly configuring complex applications[EB/OL]. 2019. https://github.com/facebookresearch/hydra.

[78] BIEWALD L. Experiment Tracking with Weights and Biases[EB/OL]. 2020. https://www.wandb.com/.

[79] HE K, ZHANG X, REN S, et al. Deep Residual Learning for Image Recognition[C/OL]//2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016. [S.l.]: IEEE Computer Society, 2016: 770-778. https://doi.org/10.1109/CVPR.2016.90. DOI: 10.1109/CVPR.2016.90.

[80] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale[C/OL]//9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021. [S.l.]: OpenReview.net, 2021. https://openreview.net/forum?id=YicbFdNTTy.

[81] RADFORD A, KIM J W, HALLACY C, et al. Learning Transferable Visual Models From Natural Language Supervision[C/OL]//MEILA M, ZHANG T. Proceedings of Machine Learning Research: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event: vol. 139. [S.l.]: PMLR, 2021: 8748-8763. http://proceedings.mlr.press/v139/radford21a.html.

[82] KIM W, SON B, KIM I. ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision[C/OL]//MEILA M, ZHANG T. Proceedings of Machine Learning Research: Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event: vol. 139. [S.l.]: PMLR, 2021: 5583-5594. http://proceedings.mlr.press/v139/kim21k.html.

# 致　谢

## .1　毕设涉及的论文发表

本论文包含两篇第一作者论文，第一篇为 ACL 2022（CCF-A）主会论文，第二篇投稿至 EMNLP 2022（CCF-B）。

[1] **Xichen Pan**, Peiyu chen, Yichen Gong, Helong Zhou, Xinbing Wang and Zhouhan Lin. Leveraging Unimodal Self-Supervised Learning for Multimodal Audio-Visual Speech Recognition[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022:4491-4503.

[2] **Xichen Pan**, Zekai Li, Yichen Gong, Helong Zhou, Xinbing Wang and Zhouhan Lin. Integrating Word Boundaries for Word Level Lipreading[C]//EMNLP 2022 submission.

## .2　致谢

四年匆匆，值此尾声；只言片语，谨表谢意。

首先我想感谢我的指导老师林洲汉，他是我科研上的启蒙者和领路人。我现在还能回忆起 2021 年 2 月 1 日下午给林老师发出的第一封邮件。作为他在交大第一个进组实习的学生，我得到了老师太多的指导和帮助。他言传身教，无论是高层次的对待科研的态度，思考问题的角度，解决问题的思路，亦或是低层次的如何看待某个观点，如何理解某个想法，如何优化工程实现，我都从中收获良多。在研究之外，我也获得了林老师在生活和未来发展上的关心，尤其感谢林老师在我博士选校中非常有帮助的建议。

我还想感谢地平线实习期间的导师宫一尘。在我实习的一年多时间里，他在科研和生活上都给了我极大的帮助，提供了宝贵的建议和讨论。感谢宫老师和王子扬、罗述杰、迟旭然、王琳、钱劲翔等同事提供的工业界方案的分享，这帮助我认识到了学界和工业界解决问题思路的差异。我还要感谢地平线多模语音算法组这个高效、轻松、有趣的部门，让我拥有了愉快的实习体验。感谢地平线的周贺龙老师，他在自监督学习方面分享了很多前沿的想法和理解。

感谢 LUMIA 实验室和计算机系中的薛昊天、陈沛宇、黎泽楷、徐昕宇等同学，他们为我的研究提供了很多讨论和建议。

在研究之外，我想感谢在申请季中作为我的推荐人的林洲汉、宫一尘、杨旸三位老师；感谢我的两位室友薛昊天和陈沛宇在学习生活上对我的支持与陪伴，我们一起度过了本科四年时光；感谢黄向鸿、沈显星、万懿、林圣凯、林航、黄首杰、姚杰腾、汤学涵、吴仕渠、姚迪熙、朱展达等诸多好友在我本科期间的关心和帮助；感谢 F1802018 班、F1803304 班的全体同学。同窗四年，情谊良多，愿诸君跃入人海，各自精彩。

感谢母校上海交通大学对我的培养，交大为我提供了一个能够接触前沿研究的高水准平台。特别感谢交大的教学老师、行政人员和后勤队伍，你们的工作让交大成为一个开放高效，活力十足，任由学生自由施展的平台，这里想特别感谢计算机系任庆生老师、张同珍老师、龙环老师，以及教务办魏冬鹤老师、思政肖汉老师、毕设易冉老师、东 13 楼李莉阿姨。

最重要的是感谢我的父母潘强先生和林宏女士，在我生命中永远支持着我，守护着我。

感谢SJTUThesis制作的 LaTeX 模板。衷心感谢参与本论文评审和答辩的老师们。

# MULTIMODAL AUDIO-VISUAL SPEECH RECOGNITION SYSTEM BASED ON PRE-TRAINED MODELS

Multimodal audio-visual Speech Recognition is a speech recognition task that leverages both an audio input of human voice and an aligned visual input of lip motions. Unlike auto speech recognition, which recognizes only audio input into speech content, audio-visual speech recognition has better recognition robustness under a noisy environment due to the involvement of visual modality. It has been one of the successful application fields that involve multiple modalities in recent years. Due to the limited amount of labeled, multimodal aligned data and the difficulty of recognition from the visual inputs (i.e., lip-reading), it is a challenging task to tackle.

Currently, popular modern audio-visual speech recognition models use temporal modeling modules based on attention mechanisms. Training such models places higher demands on the size of labeled and aligned multimodal training data. Due to the high cost of collecting such data, it makes a lot of sense to make use of unlabelled unimodal data.

Self-supervised learning is an intuitive way of leveraging these large-scale unlabeled unimodal data to improve the performance of multimodal audio-visual speech recognition. Because it is able to learn the general representation of data through simple tasks that do not require labeling. Contrastive learning has become the most impactful learning scheme in this field. In audio speech processing, contrastive predictive coding has shown its effectiveness in speech recognition. In the vision domain, models using contrastive learning for general representation learning have been proven successful, with outstanding performances on downstream tasks such as classification and detection. However, how to successfully apply unimodal self-supervised learning to multimodal audio-visual speech recognition faces two challenges:

(1) how to apply the pre-trained model to visual speech recognition

In the audio modality, self-supervised pre-training models such as wav2vec[27], vq-wav2vec[28], wav2vec 2.0[22], WavLM[29] and HuBERT[30] have demonstrated outstanding auto speech recognition performance. However, there are no self-supervised pre-training models in the visual modality that can be used for visual speech recognition. However, for general-purpose visual representation extractors, models such as MoCo[23], MoCo v2[8], SimCLR[24], SimSiam[31] and BYOL[25] have shown their competitive performances on tasks such as classification and detection. There are some difficulties in using such models as a visual feature extractor. Because the lip motions between frames are very important in audio-visual speech recognition while the images used for pre-training are mostly common objects, it is still unknown whether pre-trained visual models designed for tasks with single-frame images can be applied to audio-visual speech recognition.

(2) how to integrate unimodal pre-trained models into a multimodal scenario

The use of self-supervised learning on multimodal audio-visual speech tasks has not been adequately explored. Shukla et al.[26] is among the few attempts in this facet in which it predicts lip motions from audio inputs. Their proposed learning schemes yield strong emotion recognition re-

sults but are relatively weak in speech recognition. It is unknown whether the two feature extractors inherited from unimodal self-supervised learning can cooperate well and achieve competitive results after being fine-tuned on multimodal data.

For the first problem, we replace the 2-D convolution in the generic visual self-supervised pre-trained model with a randomly initialized 3-D convolution, enabling the pre-trained feature extractor to capture temporal correlations between visual frames as well as reducing the impact of domain shift between pre-training and training data. We test the proposed method on a word-level visual speech recognition task and show that the modified pre-training model effectively improves recognition performance.

We also further explore the word-level visual speech recognition task. Since the attention mechanism-based models are not sensitive to the position of input, for short sequence and small data scale tasks, they are not as good as recurrent neural network and convolutional neural network models that have a strong prior on position. However, they are more suitable for long sequence and large data scale tasks. Moreover, the models based on the attention mechanism cannot effectively utilize word boundaries information, which leads to its poor performance on word-level visual speech recognition. We use attention masks to leverage the word boundaries information. Frames within the word boundaries can see all frames thus are able to integrate the environment and context information in the frames outside the word boundaries, we use a special `[CLS]` embedding which can only see target word frames to integrate their information. This enables the attention mechanism-based models to effectively use word boundaries information to improve the performance of word-level visual speech recognition. We also use the viseme-level CTC loss to give an additional supervised signal to the decoder output, which can help the model optimize the features of the target word frame. Finally, the proposed model achieves a top-1 classification accuracy of 89.1% on the widely accepted LRW (Lip Reading in the Wild) dataset, with an absolute improvement of 0.6% over the current state-of-the-art model.

For the second problem, we design a multimodal audio-visual speech recognition framework that recognizes aligned audio-visual inputs into character-level outputs by combining CTC and seq2seq decoding. We inherit partial parameters of pre-trained audio and visual front-ends into this framework and obtain the final audio-visual model through a training pipeline.

Our experiments show that the two front-ends inherited from unimodal self-supervised learning can be efficiently trained through the pipeline, and the proposed model can get competitive recognition results through fine-tuning. Even without an external language model, the proposed model achieved a WER (word error rate) of 2.6%, raising the state-of-the-art performances on the widely accepted LRS2 (Lip Reading Sentences 2) dataset by a relative improvement of 30%. Compared to the baseline model without pre-trained front-ends, the proposed model has better recognition robustness under all SNR noise levels. We also test models' recognition performance under low resources, and experiments demonstrate that for the audio-only setting, the proposed model can achieve a WER of 3.4% using only 28 hours of training data, which is even better than the current state-of-the-art model. However, for the visual-only setting, our model still cannot achieve a competitive performance with only 28 hours of training data. This could be due to the small size of the visual modality's pre-training data, we believe that pre-training with larger-scale face data will result in even more improvement. Meanwhile, video-level pre-training models can be further introduced

to improve the ability to model the temporal correlation between frames and thus improve visual speech recognition.

We code a multimodal audio-visual speech recognition toolkit based on the above two models' implementations. Especially, we build preprocessing and data pipelines, organize predefined modules and utilities, and provide a training code framework using Pytorch Lightning + Hydra + WandB. We also provide our model implementation examples using the above-mentioned toolkit. We believe by using the toolkit, researchers can reproduce our results and implement their own models more easily.

In general, in this thesis we explore the application of unimodal self-supervised pre-training on multimodal audio-visual speech recognition and demonstrate its effectiveness. The proposed word-level visual speech recognition model obtains state-of-the-art performance on the LRW dataset, and the proposed sentence-level audio-visual speech recognition model raises the state-of-the-art performance on the LRS2 dataset by a large margin.