多尺度门控时空增强的唇语识别方法

马金林[1,2)], 郭兆伟[1)*], 马自萍[3)], 吕鑫[1)]

[1)] (北方民族大学计算机科学与工程学院 银川 750021)
[2)] (北方民族大学图像图形智能处理国家民委重点实验室 银川 750021)
[3)] (北方民族大学数学与信息科学学院 银川 750021)
(20217417@stu.nmu.edu.cn)

摘 要: 针对唇语识别模型中的普通卷积对唇部形变缺乏鲁棒性和不能有效地提取时间信息的问题, 提出时空增强与多尺度时间卷积网络(MSTCN)结合的唇语识别方法. 首先设计沙漏型卷积块(FCB), 增强网络对唇部形变的鲁棒性; 然后使用门控时移融合(GSF)模块提高前端网络的时间信息提取能力; 再基于 FCB 和 GSF 设计混合 3D 和 2D 卷积的时空增强网络 STABNet; 最后将 STABNet 作为前端网络, MSTCN 作为后端网络, 设计唇语识别模型. 在 LRW 数据集上的实验结果表明, 与基线模型相比, 所提方法的准确率提升 4.15 个百分点, 达到 89.45%, 而模型的参数量仅增加 3.17M. 在 GRID 数据集上准确率达到 97.45%, 超过大部分对比模型.

关键词: 唇语识别; 多尺度时间卷积网络; 时空增强; 门控时移融合模块
中图法分类号: TP391 **DOI:** 10.3724/SP.J.1089.2023-00478

Multi-Scale Gated Spatio-Temporal Enhancement for Lip Recognition

Ma Jinlin[1,2)], Guo Zhaowei[1)*], Ma Ziping[3)], and Lyu Xin[1)]

[1)] (*College of Computer Science and Engineering, North Minzu University, Yinchuan* 750021)
[2)] (*Key Laboratory of Image and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan* 750021)
[3)] (*College of Mathematics and Information Science, North Minzu University, Yinchuan* 750021)

Abstract: In order to solve the problem that the ordinary convolution is not robust to lip deformation and the accuracy of lip recognition is not high because it cannot effectively extract temporal information, a lip recognition method based on spatiotemporal enhancement network and multi-scale temporal convolution network is proposed in this paper. Firstly, an hourglass convolution block (FCB) is designed to enhance the robustness of the network to lip deformation. Secondly, a gate-shift-fuse module (GSF) is used to improve the ability of extracting time information from the front-end network. Then, based on FCB and GSF, a hybrid 3D and 2D convolutional spatio-temporal augmented block network STABNet is designed. Finally, using STABNet as the front-end network and multi-scale time convolution (MSTCN) as the back-end network, the lip recognition model is designed. Experiments show that the proposed method outperforms the baseline model by 4.15 percentage points on the LRW dataset, reaching 89.45%, and the number of model parameters increases by only 3.17M. An accuracy of 97.45% was achieved on the GRID dataset, surpassing the performance of most existing models.

Key words: lip recognition; multi-scale temporal convolution network; spatio-temporal enhancement; gate-shift-fuse module

收稿日期: 2023-07-26; 修回日期: 2025-03-04. **基金项目**: 国家自然科学基金(62462001); 宁夏自然科学基金(2024AAC03147); 中央高校基本科研业务费专项资金(2023ZRLG02). **马金林**(1976—), 男, 博士, 副教授, 博士生导师, CCF 会员, 主要研究方向为计算机视觉、唇语识别; **郭兆伟**(1997—), 男, 硕士, 论文通信作者, 主要研究方向为唇语识别; **马自萍**(1977—), 女, 博士, 教授, 博士生导师, 主要研究方向为医学图像处理、计算机视觉; **吕鑫**(1998—), 女, 硕士, 主要研究方向为唇语识别.

现有的唇语识别模型通常采用2D卷积神经网络(convolutional neural network, CNN)[1-2]或者2D CNN与3D CNN结合[3]的方式提取唇部图像的时空特征. 使用2D CNN的方式有效地捕获了单帧唇部图像的特征; 2D CNN与3D CNN结合的方式可以提取部分唇部图像序列间的时间信息, 进一步提高模型的准确率. 然而, 这类方法存在2个问题: (1) 纯2D CNN结构没有考虑整个图像序列的时空信息, 对连续帧的处理能力不足, 丢失帧间的时间相关性, 严重降低了唇语识别模型的性能[4]; 混合2D CNN和3D CNN结构虽然缓解了时序信息的不足, 但是浅层的3D CNN仍然无法捕获足够时序信息, 限制了模型的性能. (2) 这些方法都忽略了唇语数据集的数据增强(如水平翻转)的作用, 由于普通卷积对形变图像的特征不具有不变性, 使得模型对形变图像缺乏鲁棒性, 因此进一步减弱了模型的特征提取能力.

本文提出时空增强网络(spatial-temporal-augment block network, STABNet)结合MSTCN (multi-scale TCN)的唇语识别方法, 其由前端网络STABNet和后端网络MSTCN这2部分组成: STABNet结合3D CNN与具有提取时序特征能力的2D CNN, 用于提取唇语序列图像的时空特征; 为了对STABNet捕获的特征进行时序建模, 选用速度快、参数量小的时间卷积网络(temporal convolutional network, TCN)的变种网络——MSTCN.

本文方法首先针对时间信息提取不充分的问题, 引入可忽略参数量的门控时移融合(gate-shift-fuse, GSF)模块, 使模型具有更好的时序特征提取能力; 然后针对普通卷积对形变图像的鲁棒性不足的问题, 提出沙漏型卷积块(funnel-type convolutional block, FCB), 以提高模型的鲁棒性; 最后在LRW和GRID数据集上进行对比实验, 以验证其有效性.

1　相关工作

1.1　ResNet

ResNet[5]是一种深度残差网络, 自提出以来受到了广泛的关注, 其核心思想是通过在网络中添加跨层连接, 解决深度神经网络中的梯度消失和梯度爆炸问题, 以利于特征的传播. Weng等[6]采用灰度视频流和光流设计唇语识别方法, 并以ResNet-34作为特征提取网络; Martinez等[7]在前端网络中使用ResNet-18提取视频特征. 目前, ResNet及其变种网络已经成为唇语识别研究中的主流特征提取网络.

针对ResNet改进的模型有ResNeXt[8]和iResNet[9]等. 其中, 在网络深度相同时, iResNet不仅未增加计算量, 而且具有更高的识别精度, 并已在图像分类、目标检测和视频动作识别研究中验证了模型的有效性, 其最深的网络超过3 000层, 而大多数网络在这种深度难以收敛. iResNet有以下特点: (1) 构造新的网络结构, 增强了信息流动和特征表达能力; (2) 在几乎不增加计算量的情况下, 增强了残差模块的学习能力; (3) 相同深度时的网络性能更好.

如图1b所示, iResNet采用一种快捷映射方式分别处理空间维度和通道维度, 首先利用步长为2的3×3卷积进行最大池化, 实现空间维度的降采样; 然后使用步长为1的1×1卷积进行通道维度的变换. 与图1a所示的ResNet快捷映射方式相比, 最大池化操作能够有效地提取特征图的显著信息, 避免信息丢失.

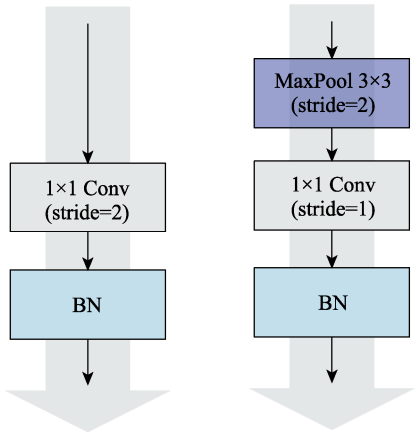


图1　快捷映射方式

1.2　非对称卷积

Szegedy等[10]提出Inception-v3, 将普通卷积分解为更小的卷积, 证明了非对称卷积结构可以有效地降低模型参数量, 并节省计算资源. Ding等[11]提出ACNet, 使用非对称卷积增强普通卷积的表征能力, 并将非对称卷积与普通卷积融合, 在不增加额外推理时间的情况下提高了卷积的特征提取能力. ACNet由非对称卷积模块(asymmetric convolution block, ACB)构建, 使用并行的1×3, 3×1和3×3卷积替代3×3卷积. ACB的原理如图2所示, 利用大小兼容的卷积核代替单个卷积, 得到的输出按照对应位置求和, 提高模型提取特征的能力.

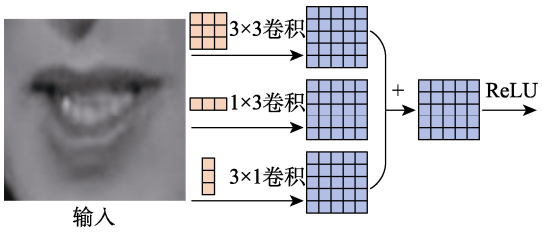


图 2 ACB 的原理

1.3 时空特征提取

唇语识别模型以包含唇部运动变化的视频帧作为基础数据，唇部的时序变化信息是识别模型关注的重点. 虽然3D CNN已经被证明可以有效地提取时间序列信息，但由于其对硬件资源的要求较高，限制了自身的发展.

近年来，越来越多的研究者尝试将具有时序特征提取能力的模块插入 2D CNN，并以此进行视频动作识别和视频处理. Lin 等[12]提出时间移位模块(temporal shift module, TSM)，在时间维度上向前和向后移动部分通道，以加强相邻帧的信息交互；TSM 插入 2D CNN 后，可以在保持 2D 网络复杂度的情况下实现时间建模. TSM 的原理如图 3 所示.

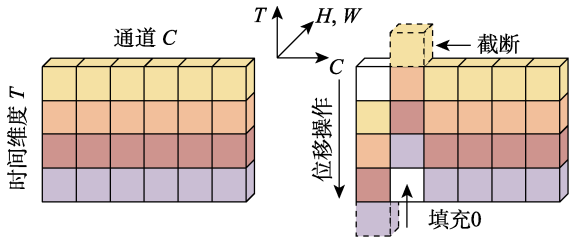


图 3 TSM 的原理

Hao 等[13]将 TSM 引入唇语识别后取得了较好的识别性能，但 TSM 缺乏对动作的显式时间建模，导致其模型无法充分地捕获运动信息；Wang 等[14]提出 ACTION-Net，设计时空、通道和运动激励的动作模块，并将其嵌入 ResNet 后应用于动作识别任务；Sudhakaran 等[15]基于 TSM 提出 GSF 模块，利用分组空间门控，时间移位和特征融合提取时空特征，并与 2D 网络结合构建了高效的时空特征提取器，该模块不仅具有与 3D 模型相似的时序特征提取能力，而且保持与 2D 模型相当的参数量和计算成本.

2 本文方法

本文方法由 STABNet 和 MSTCN 组成，其结构如图 4 所示.

STABNet 由 3D 卷积和 4 个堆叠的 STAB 组成. 3D 卷积使用 5×7×7 的卷积核，后接批归一化(batch normalization, BN)和 ReLU 激活函数. STAB 由 FCB 和 GSF 顺序连接组成，其中，FCB 由非对称卷积与常规卷积构成，用于提高模型鲁棒性；GSF 由分组门控、前向后向时间偏移和特征融合模块构成，用于增强模型时序信息提取能力.

MSTCN 处理不同尺度的时间信息，使模型具有更强的时序建模能力. MSTCN 的卷积核尺寸分别为 3, 5 和 7.

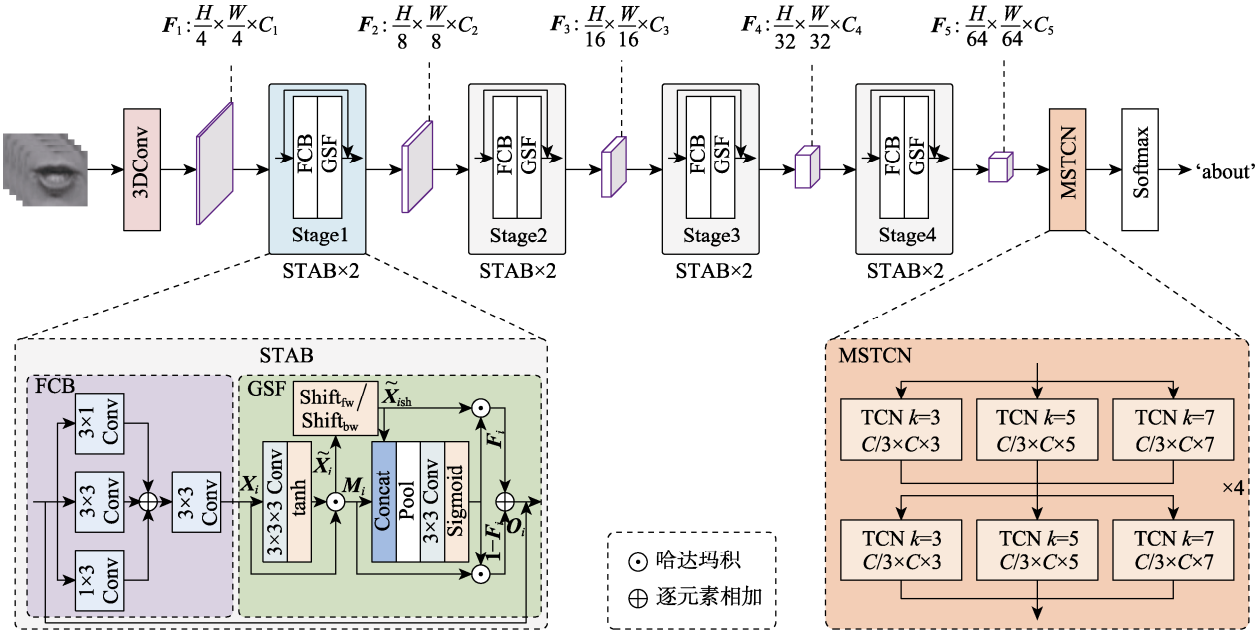


图 4 本文方法结构

2.1 STABNet

2.1.1 FCB

主流的唇语识别模型是 3D CNN 和 2D CNN 的混合网络. 由于 2D 网络普遍采用普通卷积提取特征, 因此模型对唇部形变缺乏鲁棒性. 针对此问题, 本文设计了 FCB, 其结构如图 5 所示. 首先将非对称卷积(1×3 和 3×1)与卷积(3×3)混合构成 FCB, 采用这种大小“兼容”的 2D 卷积得到与输入图像相同分辨率的输出; 然后将该输出按对应位置求和, 融合各卷积的特征图; 最后在每个卷积层使用 BN 操作加快收敛速度.

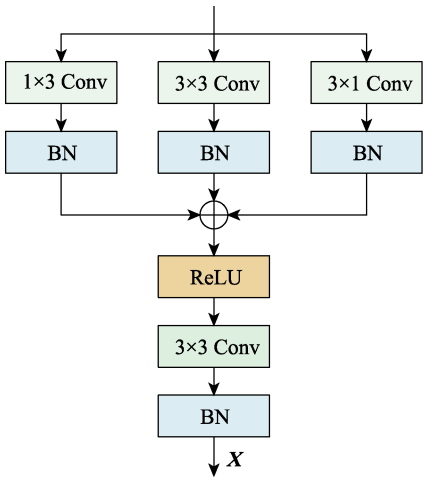


图 5　FCB 的结构

FCB 有 3 个优点: (1) 采用 1×3 和 3×1 卷积的参数量是 3×3 卷积的 1/3, 整体结构只增加少量参数; (2) 采用并行的方式, 同时进行多个卷积运算, 加速卷积过程; (3) 与仅使用 3×3 卷积相比, 水平卷积(1×3)对上下翻转有更强的鲁棒性和特征提取能力. 如图 6 所示, 1×3 卷积在图像上下翻转后得到的图像与原始输入相同, 而 3×3 卷积却无法得到.

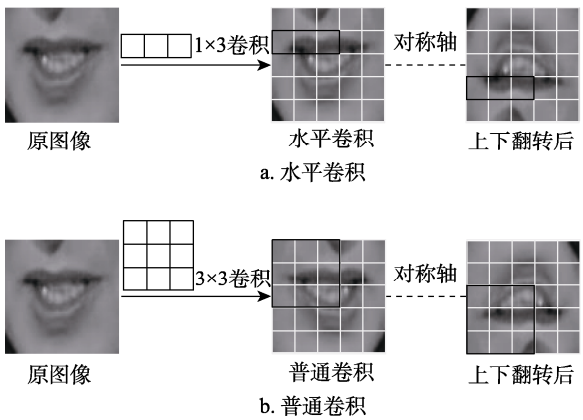


图 6　水平卷积与普通卷积示意图

FCB 采用简单的 3×3 卷积融合特征整合不同方向和尺度的特征, 获得更丰富和抽象的特征. 另外, 连续的 3×3 卷积可以增大感受野, 使后续层的神经元接触更大的输入区域, 并使网络更好地理解输入数据的复杂结构和模式.

通过以上设计, 模型可以学习更加多样化的特征表示, 有益于增强模型对唇部形变的鲁棒性, 提高对不可见数据的泛化能力.

2.1.2 GSF

在唇语识别中, 有效地利用帧间的时间信息可以提高模型的性能, 然而, 大多数模型使用的单层 3D CNN 不能充分地提取时间序列信息. 为此, 本文引入 GSF 模块[15], 利用分组空间门控分解输入张量和通道加权来融合分解张量. 将 GSF 模块嵌入 2D CNN, 可使网络保持 2D CNN 的低复杂度, 同时具有捕获空间和时间特征的能力. 如图 7 所示, GSF 模块由分组空间门控(Gate)操作, 时间移位(Shift)操作和融合(Fuse)操作组成, GSF 模块将输入张量按照通道分解为 2 组, 每组分别进行处理. Gate 操作控制组之间的交互, 通道加权机制融合分解后的张量, 可获得丰富的时空信息. Gate 操作依据时间自适应地路由特征, 并以数据依赖的方式进行组合, 使 GSF 模块能够捕获更多的时序信息. 因此, GSF 模块可以嵌入 2D 网络, 使其具有捕获时间和空间特征的能力.

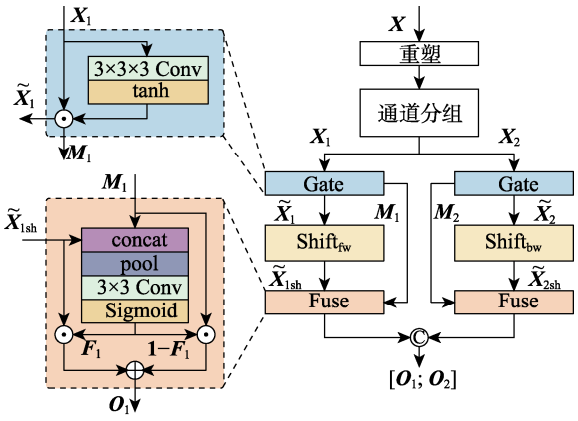


图 7　GSF 的结构

FCB 的输出 $\boldsymbol{X}$ 作为 GSF 的输入, 将其变为 $C\times T\times H\times W$ 的张量; 其中, C,T,H,W 分别代表通道、时间、特征图的高和宽 4 个维度. $\boldsymbol{X}$ 按通道维度分为 $\boldsymbol{X}_1$ 和 $\boldsymbol{X}_2$ 这 2 组, 分别输入 GSF 模块. 该过程表示为

$$\begin{cases}\tilde{\boldsymbol{X}}_1=\tanh\left(\boldsymbol{W}_{g1}*\boldsymbol{X}_1\right)\odot\boldsymbol{X}_1\\ \tilde{\boldsymbol{X}}_2=\tanh\left(\boldsymbol{W}_{g2}*\boldsymbol{X}_2\right)\odot\boldsymbol{X}_2\end{cases}\tag{1}$$

$$\begin{cases} \boldsymbol{M}_1 = \boldsymbol{X}_1 - \tilde{\boldsymbol{X}}_1 \\ \boldsymbol{M}_2 = \boldsymbol{X}_2 - \tilde{\boldsymbol{X}}_2 \end{cases} \tag{2}$$

$$\begin{cases} \tilde{\boldsymbol{X}}_{1\mathrm{sh}} = \mathrm{Shift}_{\mathrm{fw}}\left(\tilde{\boldsymbol{X}}_1\right) \\ \tilde{\boldsymbol{X}}_{2\mathrm{sh}} = \mathrm{Shift}_{\mathrm{bw}}\left(\tilde{\boldsymbol{X}}_2\right) \end{cases} \tag{3}$$

$$\begin{cases} \boldsymbol{F}_1 = \sigma\left(\boldsymbol{W}_{\mathrm{f}1}\left[\theta\left(\tilde{\boldsymbol{X}}_{1\mathrm{sh}}; \boldsymbol{M}_1\right)\right]\right) \\ \boldsymbol{F}_2 = \sigma\left(\boldsymbol{W}_{\mathrm{f}2}\left[\theta\left(\tilde{\boldsymbol{X}}_{2\mathrm{sh}}; \boldsymbol{M}_2\right)\right]\right) \end{cases} \tag{4}$$

$$\begin{cases} \boldsymbol{O}_1 = \boldsymbol{F}_1 \odot \tilde{\boldsymbol{X}}_{1\mathrm{sh}} + \left(\boldsymbol{1} - \boldsymbol{F}_1\right) \odot \boldsymbol{M}_1 \\ \boldsymbol{O}_2 = \boldsymbol{F}_2 \odot \tilde{\boldsymbol{X}}_{2\mathrm{sh}} + \left(\boldsymbol{1} - \boldsymbol{F}_2\right) \odot \boldsymbol{M}_2 \end{cases} \tag{5}$$

其中，$\boldsymbol{W}_{\mathrm{g}1}$ 表示 GSF 模块左卷积的权重，$\boldsymbol{W}_{\mathrm{g}2}$ 表示 GSF 模块右卷积的权重；$\boldsymbol{W}_{\mathrm{f}1}$ 表示 $\boldsymbol{X}_1$ 融合时卷积权重，$\boldsymbol{W}_{\mathrm{f}2}$ 表示 $\boldsymbol{X}_2$ 融合时卷积权重；$\odot$ 表示哈达玛积；$\mathrm{Shift}_{\mathrm{fw}}$ 表示前向 Shift，$\mathrm{Shift}_{\mathrm{bw}}$ 表示后向 Shift；$(\cdot;\cdot)$ 表示拼接；σ 表示 Sigmoid；θ 表示空间平均池化操作. 以 $\boldsymbol{X}_1$ 为例，计算 GSF 的过程如下：

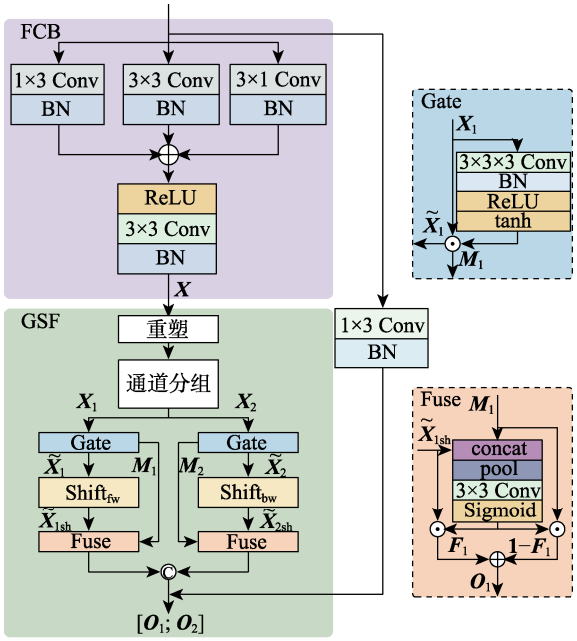
(1) $\boldsymbol{X}_1$ 输入经过 Gate 操作，得到一对门控特征及其差分特征. 式(1)中，经 3×3×3 卷积，非线性激活函数 tanh 与 $\boldsymbol{X}_1$ 进行哈达玛积得到 $1\times T\times H\times W$ 的门控特征 $\tilde{\boldsymbol{X}}_1$，通过式(2)得到差分特征 $\boldsymbol{M}_1$，捕获上下文差异特征. tanh 起到 Gate 作用，其取值范围为 (−1, +1). 当某个特征位置门控值为 0，且 Shift 操作后特征为+1 时，在该位置执行加权时间特征平均；如果 Shift 操作后的特征值为−1 时，在该位置执行加权时间特征差. 使用 tanh 门控机制自适应地选择这 2 种运算模式.

(2) 如图 7 所示，采用 TSM[12]的双向移位方式融合时序信息. 式(3)中，$\tilde{\boldsymbol{X}}_1$ 按照时间维度进行前向移位操作($\tilde{\boldsymbol{X}}_2$ 进行后向移位操作)，注入时间信息得到移位特征 $\tilde{\boldsymbol{X}}_{1\mathrm{sh}}$，其形状为 $C\times T\times 1\times 1$.

(3) 将得到的移位特征 $\tilde{\boldsymbol{X}}_{1\mathrm{sh}}$ 与其对应的差分特征 $\boldsymbol{M}_1$ 经过图 7 的 Fuse 操作进行特征融合，获得更抽象的时空特征. 式(4)中，$\tilde{\boldsymbol{X}}_{1\mathrm{sh}}$ 与 $\boldsymbol{M}_1$ 拼接，经过空间平均池化得到形状为 $2\times C\times T$ 的特征，通过 3×3 卷积和 Sigmoid 操作得到形状为 $1\times C\times T$ 的权重矩阵 $\boldsymbol{F}_1$；然后将 $\boldsymbol{F}_1$ 的形状重塑为 $C\times T\times 1\times 1$. 式(5)中，与 Shift 操作后的特征及其特征差值进行通道加权平均得到融合特征 $\boldsymbol{O}_1$，当一个特征通道的融合权重为 0 时，仅保留具有空间信息的 $\boldsymbol{M}_1$ 特征；如果融合权重为 1，舍弃 $\boldsymbol{M}_1$ 空间特征，并传播 Shift 操作后的时序特征；对于其他权重值，通过混合通道有选择地混合这 2 组特征. 采用这种方式可保留各个通道捕获的信息，增强模型的特征表示能力.

2.1.3 STAB

STAB 由 FCB 和 GSF 模块组成，其结构如图 8 所示. 其中，FCB 利用非对称卷积和普通卷积提取空间特征，每个卷积层都使用 BN 操作加快收敛速度，可有效地提高模型对形变图像的鲁棒性；GSF 模块通过分组 Gate，Shift 和 Fuse 操作将时序信息注入 2D 网络，其参数量可忽略不计，应用该模块可增强模型提取唇部序列间时序信息的能力. FCB 得到的形状为 $C\times H\times W$ 的张量经 GSF 模块重塑为 $C\times T\times H\times W$，按通道分为 2 组输入 Gate 单元，得到 2 组 Gate 特征及其对应的差分特征. Gate 特征用于 Shift 操作得到时间信息，与对应的差分特征拼接后进行 Fuse 操作.



⊙ 哈达玛积；⊕ 逐元素相加；© 拼接

图 8 STAB 的结构

综上所述，STAB 基于 2D CNN 结构提取空间特征和时间特征，因此，STABNet 由 STAB 堆叠而成.

2.2 MSTCN

与循环神经网络 LSTM 和 GRU 相比，TCN 通过并行计算处理时间序列数据，每个卷积核独立地处理特征图的不同位置，大大减少了计算时间，因此其在计算效率上具有优势. 如图 4 所示，本文堆叠 4 层 MSTCN 实现序列建模，每层 MSTCN 使用 3 种不同尺寸的卷积核(3，5 和 7)捕捉不同时间尺度的特征. 其中，较小的卷积核可以捕捉短期依赖关系，而较大的卷积核可以捕捉长期依赖关系，

使 TCN 能够更好地提取局部特征. 由于 TCN 是基于局部窗口计算的, 因此每个独立窗口的梯度可以直接传播到更早的时间步, 与循环神经网络相比, TCN 不会出现梯度消失问题.

3　实验及结果分析

3.1　数据集和数据预处理

本文使用唇语数据集 LRW[1]和 GRID[16]进行实验. 其中, LRW 数据集包含 500 个单词, 每个单词长度为 5~10 字符, 由 800~1 000 个视频片段构成训练集, 50 个片段的验证集和测试集, 总时长 173 h. 由于该数据集中说话人的头部姿势和灯光条件具有明显变化, 非常接近真实场景下的任务状态, 因此比其他词级数据集更具挑战性. GRID 数据集包含 34 位说话人的视频片段, 每位说话人有 1 000 个语句, 总持续时间 28 h; 每个样本是一位说话人的 3 s 视频, 帧率为 25 帧/s. 该数据集具有固定语法结构, 由“命令(4)+颜色(4)+介词(4)+字母(26)+数字(10)+副词(4)”组成, 它包含 51 个不同的词, 每个句子都是这些词的随机组合.

LRW 数据集使用 Dlib 工具[17]对每个视频序列的人脸进行关键点检测和定位, 其预处理过程如下: 首先通过关键点定位找到唇部区域, 并裁剪为 96×96 像素的图像; 然后将图像随机裁剪为大小 88×88 像素的唇部图像序列, 对图像进行灰度化处理, 以 0.5 的概率随机水平翻转; 最后对图像进行归一化处理后输入模型. 为了提高模型的鲁棒性, 采用 Mixup[18]的数据增强技术将 2 个不同的样本及对应的标签 (x_1, y_1) 和 (x_2, y_2) 进行线性组合, 生成新样本 x' 和新标签 y', 公式为

$$x' = \lambda x_1 + (1-\lambda) x_2,$$

$$y' = \lambda y_1 + (1-\lambda) y_2.$$

其中, λ 是由 α 和 β 的贝塔分布计算出来的混合系数, $\lambda = B(\alpha, \beta)$, 本文中, α 和 β 的取值均为 0.4.

GRID 数据集的处理方式与 LRW 数据集基本相同, 区别在于输入唇部图像序列大小为 160×80 像素. 因为 GRID 数据集中每个视频包含 6 个单词, 所以本文根据 LipNet[19]提供的视频注释文件, 将每个视频中不同词图像序列处理为与 LRW 数据集相同的格式作为输入, 并采用相同的数据增强方法.

3.2　实验平台和参数设置

本文实验基于 PyTorch 1.8.1 框架, 在显存为 32 GB 的 NVIDIA-Tesla V100 单张 GPU 上训练, 批次量为 16, 轮次为 100. 使用交叉熵损失函数和 Adamw 优化算法, 权重衰减率为 1×10^{-4}, 初始学习率为 3×10^{-4}, 采用余弦调度器调整学习率, 公式为

$$\eta_t = \frac{1}{2}\eta_0\left(1+\cos\left(\frac{\pi t}{T}\right)\right).$$

其中, η_0 表示初始学习率; t 表示批次数; T 表示总批次数.

3.3　实验设计

首先, 在 LRW 数据集上确定模型结构并验证模型性能; 其次, 在 GRID 数据集上验证模型的泛化能力.

(1) STABNet 的结构选取实验. 验证 STABNet 中各子结构的有效性, 确定该网络的结构.

(2) MSTCN 选取实验. 考察 MSTCN、双向模块循环单元(bi-gated recurrent units, BiGRU)和 TCN 对模型识别性能的影响.

(3) GSF 嵌入位置和数量的选取实验. 探究 GSF 模块的嵌入位置和数量对模型性能的影响.

(4) 在不同数据集上与主流方法的识别性能对比.

(5) 在 LRW 数据集上, 困难样本的识别能力对比.

3.4　STABNet 对识别性能的影响

为了探究 STABNet 中 2D 网络对识别性能的影响, 本文在 LRW 数据集上进行了充分的实验, 结果如表 1 所示. 其中, 将文献[7]作为基线模型; 模型 A 在基线基础上加入 FCB; 模型 B 在模型 A 基础上增加 shortcut; 模型 C 在模型 A 基础上增加 GSF; 模型 D 在基线基础上同时引入 shortcut 和 GSF; 模型 E 在基线基础上整合 shortcut, FCB 和 GSF. “shortcut”表示每阶段间使用 iResNet[9]的快捷映射方式, 否则, 使用原始 ResNet 的快捷映射方式; 未使用 FCB 时, STABNet 每个阶段采用与 ResNet 基础块相同的卷积结构, 即 2 个 3×3 卷积堆叠.

表 1　LRW 数据集上 STABNet 中不同模块的性能对比

模型	模块			准确率/%	参数量
	shortcut	FCB	GSF		
基线	×	×	×	85.30	**36.37M**
A	×	√	×	87.50	39.53M
B	√	√	×	88.68	39.53M
C	×	√	√	87.97	39.54M
D	√	×	√	88.70	36.39M
E	√	√	√	**89.45**	39.54M

注. 粗体表示最优值.

从表 1 可以看出:

(1) 与基线模型相比, 使用不同结构的模型性能均有提升, 同时使用 3 种模块时(即模型 E)的准确率提高 4.15 个百分点.

(2) 对比模型 A 和模型 B, 未使用 shortcut 的模型取得了较差的效果, 这是因为 ResNet 结构的快捷映射是由步长为 2 的 1×1 卷积接 BN 操作构成, 该方式一定程度上引入噪声并造成信息损失; 而 iResNet 是由步长为 2, 大小为 3×3 的最大池化, 步长为 1 的 1×1 卷积和 BN 顺序连接构成, 其 1×1 卷积选择性地激活信息, 减少了信息损失.

(3) 对比模型 D 和模型 E, 当 2D 网络使用改进的 FCB 时, 与模型 D 相比, 模型 E 的准确率提升 0.75 个百分点.

本文认为, 与普通卷积相比, 非对称卷积可以增强唇部形变的鲁棒性, 提升模型的特征提取能力和表现力. 对比模型 A 和模型 C 可以看出, 当使用 GSF 模块时, 模型 C 比模型 A 的准确率提高 0.47 个百分点; 对比模型 B 和模型 E 可以看出, 不使用 GSF 模块时, 模型 B 的准确率比模型 E 低 0.77 个百分点. 分析其原因是 STABNet 的 GSF 模块使网络具有捕获并保留更多时序信息的能力.

3.5 MSTCN 对识别性能的影响

对主流后端网络对模型识别性能的影响进行实验, 包括 3 层 BiGRU, 4 层 TCN 和 4 层 MSTCN, 结果如表 2 所示. 其中, TCN 的卷积核大小为 3, MSTCN 使用大小分别为 3, 5 和 7 这 3 种不同尺度的卷积核. 可以看出, TCN 作为后端网络的模型的参数最少, 准确率仅为 86.45%, 识别性能最差; MSTCN 作为后端网络的模型取得了 89.45%的最高准确率, 参数量居中, 为 39.54M; BiGRU 作为后端网络的模型的参数量最大, 识别性能低于 MSTCN. 图 9 所示为不同后端网络的模型准确率-训练轮数的结果对比. 可以看出, 本文方法所用的 MSTCN 在每轮次的准确率都高于其他 2 种方法, 且第 97 轮的准确率相差最大.

3.6 GSF 模块位置和数量的选取

在 STABNet 基础模块 STAB 中, 对 GSF 模块

表 2 STABNet 下不同后端网络的性能对比

后端网络	准确率/%	参数量
BiGRU	88.50	62.60M
TCN	86.45	**18.17M**
MSTCN	**89.45**	39.54M

注. 粗体表示最优值.

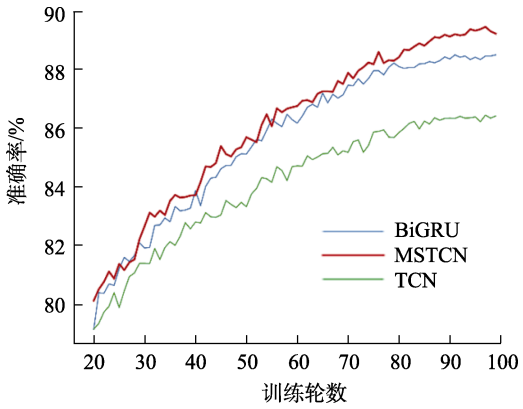


图 9 不同后端网络的模型精度对比

的位置和数量对模型性能的影响进行实验.

为了考察 GSF 模块的位置对模型性能的影响, 对多种 GSF 位置的模型性能进行实验, 嵌入位置如图 10 所示. GSF 模块的位置对识别性能的影响如表 3 所示. 其中, GSF 模块分别嵌入到基础模块的位置 *a*, 位置 *b* 和位置 *c* 处(获得前端网络 STABNet1, STABNet2 和 STABNet), 通过对比各位置的识别性能来选取最佳嵌入位置. 从表 3 可以看出, GSF 模块插入基础模块的最后位置, 即位置 *c* 处, 模型具有最佳识别性能. 图 11 所示为 GSF 的位置对模型性能的影响. 可以看出, 位置 *b* 的准确率曲线逐渐变缓, 位置 *c* 的准确率曲线稳定上升. 因此, 本文选择位置 *c* 为 GSF 模块的嵌入位置.

对多种 GSF 模块数量的模型性能进行实验. 当 GSF 数量为 4 时, 每个 stage 使用结构 1, 如图 12a

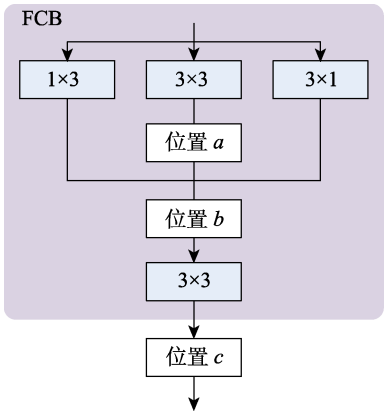


图 10 GSF 模块的嵌入位置示意图

表 3 GSF 模块位置对识别性能的影响

前端网络	位置	准确率/%	参数量
STABNet1	*a*	89.30	**39.54M**
STABNet2	*b*	88.76	**39.54M**
STABNet	*c*	**89.45**	**39.54M**

注. 粗体表示最优值.

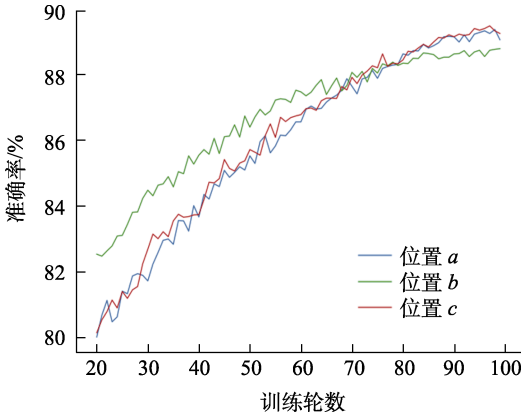


图 11　不同 GSF 模块嵌入位置下模型准确率对比

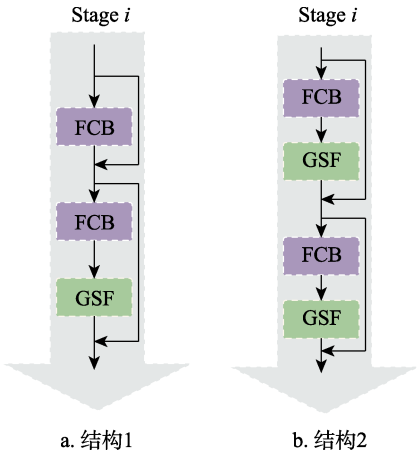


图 12　GSF 的嵌入方式

所示；而 GSF 数量为 8 时，每个 stage 使用结构 2，如图 12b 所示：这些是一种对称结构．当 GSF 数量为 6 时，stage1 和 stage3 使用结构 1，stage2 和 stage4 使用结构 2：这是一种非对称结构．GSF 模块数量对模型准确率的影响如表 4 所示．可以看出，当 GSF 数量为 8 时，识别模型获得了最好的识别效果，与 GSF 数量为 4 和 6 时相比，准确率分别提升 2.35 个百分点和 2.72 个百分点．

表 4　GSF 数量对模型准确率的影响

GSF 数量	准确率/%	参数量
4	87.10	**39.54M**
6	86.73	**39.54M**
8	**89.45**	**39.54M**

注．粗体表示最优值．

3.7　与主流方法的识别性能对比

在 LRW 和 GRID 这 2 个数据集上，对本文方法与主流方法[1,3,7,20-31]的识别性能进行实验；其中，将文献[7]作为基线模型．

在 LRW 数据集上，16 种方法的识别性能如表 5 所示．可以看出，本文方法达到 89.45%的识别准确率，比文献[7]高出 4.15 个百分点，比文献[20]高 0.75 个百分点；当前端网络为 3D Conv+ ResNet34 时，文献[21]方法使用 BiGRU 作为 MSTCN 的性能优于文献[3]的 BiLSTM；当后端网络均为 BiLSTM 时，文献[3]方法更深的 3D Conv+ResNet34 并没有取得比文献[20]更高的精度，是因为更深的网络容易过拟合；前端网络和后端网络相同时，在文献[7]和文献[22]方法中加入音频特征(multi-head visual-audio memory, MVM)使识别准确率更高，说明该特征可以提供额外信息，增强唇语特征的表达

表 5　在 LRW 数据集上 16 种方法的识别性能对比

方法	出版年	前端网络	后端网络	准确率/%
文献[1]	2017	VGGM		61.10
文献[3]	2017	3D Conv+ResNet34	BiLSTM	83.00
文献[21]	2018	3D Conv+ResNet34	BiGRU	83.39
文献[26]	2019	Multi-grained	BiConvLSTM	83.34
文献[7]	2020	3D Conv+ResNet18	MSTCN	85.30
文献[27]	2020	3D SEResNet	BiGRU	85.50
文献[23]	2021	3D Conv+ResNet18	DCTCN	88.36
文献[20]	2022	3D ResNet18	BiLSTM	88.70
文献[24]	2022	3D Conv+ResNet18	MDTCN	85.70
文献[22]	2022	3D Conv+ResNet18+MVM	MSTCN	88.50
文献[28]	2023	3D Conv+iGhost Bottneck	BiGRU+同类自知识蒸馏	87.30
文献[30]	2023	ResNet50+LPA+Transformer		84.90
文献[29]	2024			85.20
文献[31]	2024	3D Conv+DSTE+GhostNet+ME	Transformer+MS-TCN	89.21
文献[25]	2025	3D Conv+Transformer	BiGRU+weight distillation	88.30
本文(STABNet+MSTCN)		3D Conv+STABNet	MSTCN	**89.45**

注．粗体表示最优值．

能力; 当文献[7,23-24]这 3 种方法使用相同的前端网络, 后端网络为不同 TCN 变种时, 获得了较高的准确率, 说明基于TCN的后端网络可以有效地处理时序数据; 文献[25]采用知识蒸馏等方法压缩模型, 减少硬件消耗, 但模型的识别性能有所下降; 与使用 Transformer 关注长程时空特征的文献[31]相比, 本文方法获得了更高的准确率, 说明在前端关注时空特征比后端关注时空特征更有效.

在 GRID 数据集上, 8 种方法的识别性能对比如表 6 所示, 其中, 实验设置与 LRW 数据集上完全相同. 可以看出, 本文方法的准确率获得第 3 名, 比文献[30]方法低 1.45 个百分点, 比第 2 名低 1.25 个百分点, 这是因为本文方法对句子级语料未进行对齐操作, 而对比方法均使用了对齐操作, 说明本文方法对单词级语料的识别性能较好, 但对句子级语料的识别性能有待提高.

表 6 在 GRID 数据集上 8 种方法的识别性能对比

方法	年份	前端网络	后端网络	准确率/%
文献[32]	2018	3D Conv+highway network	BiGRU+cascaded attention-CTC	97.10
文献[33]	2019	3D Conv+2D CNN	BiLSTM+w-CTC	98.70
文献[34]	2019	STCNN+STPP+ResNet50	BiGRU+Hybrid/attention CTC	97.30
文献[35]	2020	3D Conv+Transformer	3D Conv+Transformer	95.50
文献[36]	2022	User-dependent Padding+CNN	GRU	92.80
文献[30]	2023	ResNet50+LPA+Transformer	ResNet50+LPA+Transformer	**98.90**
文献[37]	2024	3D Conv+EfficientNetB0	BiLSTM	96.70
本文(STABNet+MSTCN)		3D Conv+STABNet	MSTCN	97.45

注. 粗体表示最优值.

3.8 不同样本难度的识别能力对比

在不同难度样本组上, 对基线模型, 后端网络为 BiGRU(即 STABNet+BiGRU 方法), 在位置 a 嵌入 GSF 模块(即 STABNet1+MSTCN 方法)和本文方法的平均识别准确率进行实验, 验证本文方法在 LRW 数据集上对不同样本难度的识别能力. 样本难度的确定方法如下: 以基线模型对测试样本的识别准确率高低为依据, 将样本分为容易、较易、中等、困难和极难 5 组, 每组包含 100 个类别. 在 LRW 数据集上, 对 5 组难度样本的识别准确率进行实验, 结果如图 13 所示. 可以看出, 本文方法对于容易和较易这 2 组样本的识别率与基线模型相同; 但对于其他 3 组样本, 本文方法的识别准确率仍然保持在较高的水平, 尤其是对极难样本的识别超过其他 3 种方法.

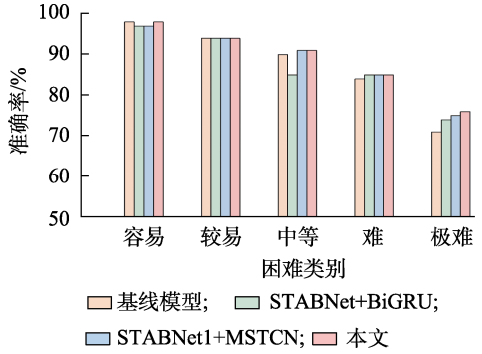


图 13 在 LRW 上 5 组难度样本的识别准确率对比

将 4 种方法对随机选取的极难样本中的 11 个单词进行识别实验, 结果如表 7 所示. 可以看出: (1) “sense”在 4 种方法中均识别错误, 可能是因为其发音口型变化小, 所有方法均没有捕获该词的独有特征. (2) 除本文方法外, 其他方法均将“sector”识别为与其发音口型相似的词“second”或“attacks”; 同样地, 其他方法均将“report”识别为发音相似的“support”或“reports”. 实验结果表明, 本文方法对区分发音口型相似的单词更有优势, 这得益于其增强了模型对于形变图像的高特征获取能力和鲁棒性.

表 7 4 种方法对 11 个极难单词的识别结果对比

标签	方法			
	基线模型	STABNet+BiGRU	STABNet1+MSTCN	本文
trying	trade	trying	trying	trying
hours	areas	areas	hours	hours
million	million	million	media	million
sense	social	businesses	state	sensor
sector	second	second	attacks	sector
report	support	support	reports	report
areas	Greece	areas	areas	areas
thing	think	tonight	thing	thing
spend	spending	spend	spend	spend
bring	pretty	pretty	bring	bring
because	build	because	parts	because

4 结　语

本文从增强唇部形变的鲁棒性和提高时间信息提取能力 2 个方面入手，提出 STABNet+MSTCN 方法. 首先提出 FCB，其有效地改善对唇部形变鲁棒性差的问题；然后采用 GSF 模块更好地捕获唇部序列间的时序特征. 实验结果证明，本文方法提高了模型的识别准确率，在不同数据集上具有较好的泛化能力. 未来的工作重点是做好语料对齐，以更好地处理句子级语料.

参考文献(References):

[1] Chung J S, Zisserman A. Lip reading in the wild[C] //Proceedings of the 13th Asian Conference on Computer Vision. Heidelberg: Springer, 2017: 87-103

[2] Zhang G Y, Lu Y Y. Research on a lip reading algorithm based on efficient-ghostNet[J]. Electronics, 2023, 12(5): Article No.1151

[3] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading[C] //Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm: ISCA, 2017: 3652-3656

[4] Sheng C C, Kuang G Y, Bai L, *et al*. Deep learning for visual speech analysis: a survey[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6001-6022

[5] He K M, Zhang X Y, Ren S Q, *et al*. Deep residual learning for image recognition[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 770-778

[6] Weng X S, Kitani K. Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading[C] //Proceedings of the 30th British Machine Vision Conference 2019. Durham: BMVA Press, 2019: Article No.269

[7] Martinez B, Ma P C, Petridis S, *et al*. Lipreading using temporal convolutional networks[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2020: 6319- 6323

[8] Xie S N, Girshick R, Dollár P, *et al*. Aggregated residual transformations for deep neural networks[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2017: 5987-5995

[9] Duta I C, Liu L, Zhu F, *et al*. Improved residual networks for image and video recognition[C] //Proceedings of the 25th International Conference on Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 9415-9422

[10] Szegedy C, Vanhoucke V, Ioffe S, *et al*. Rethinking the inception architecture for computer vision[C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2016: 2818-2826

[11] Ding X H, Guo Y C, Ding G G, *et al*. ACNet: strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 1911-1920

[12] Lin J, Gan C, Han S. TSM: temporal shift module for efficient video understanding[C] //Proceedings of the IEEE/CVF International Conference on Computer Vision. Los Alamitos: IEEE Computer Society Press, 2019: 7082-7092

[13] Hao M F, Mamut M, Yadikar N, *et al*. How to use time information effectively? Combining with time shift module for lipreading[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2021: 7988-7992

[14] Wang Z W, She Q, Smolic A. ACTION-Net: multipath excitation for action recognition[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Alamitos: IEEE Computer Society Press, 2021: 13209-13218

[15] Sudhakaran S, Escalera S, Lanz O. Gate-shift-fuse for video action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 10913-10928

[16] Cooke M, Barker J, Cunningham S, *et al*. An audio-visual corpus for speech perception and automatic speech recognition[J]. The Journal of the Acoustical Society of America, 2006, 120(5): 2421-2424

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

[18] Zhang H Y, Cisse M, Dauphin Y N, *et al*. Mixup: beyond empirical risk minimization[C] //Proceedings of the 6th International Conference on Learning Representations. Washington DC: ICLR, 2018: 1-13

[19] Assael Y M, Shillingford B, Whiteson S, *et al*. LipNet: end-to-end sentence-level lipreading[OL]. [2023-07-26]. https://arxiv.org/abs/1611.01599

[20] Ivanko D, Ryumin D, Kashevnik A, *et al*. Visual speech recognition in a driver assistance system[C] //Proceedings of the 30th European Signal Processing Conference. Los Alamitos: IEEE Computer Society Press, 2022: 1131-1135

[21] Petridis S, Stafylakis T, Ma P, *et al*. End-to-end audiovisual speech recognition[C] //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. Los Alamitos: IEEE Computer Society Press, 2018: 6548-6552

[22] Kim M, Yeo J H, Ro Y M. Distinguishing homophenes using multi-head visual-audio memory for lip reading[C] //Proceedings of the 36th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2022: 1174-1182

[23] Ma P C, Wang Y J, Shen J, *et al*. Lip-reading with densely connected temporal convolutional networks[C] //Proceedings of the IEEE Winter Conference on Applications of Computer Vision. Los Alamitos: IEEE Computer Society Press, 2021: 2856-2865

[24] Xu B Y, Wu H Y. Lip reading using multi-dilation temporal convolutional network[C] //Proceedings of the 2nd International Conference on Signal Processing and Machine Learning. Aachen: CEUR-WS.org, 2022: 50-59

[25] Wang H J, Cui B Y, Yuan Q B, *et al*. Mini-3DCvT: a lightweight lip-reading method based on 3D convolution visual transformer[J]. The Visual Computer, 2025, 41: 1957-1969

[26] Wang C H. Multi-grained spatio-temporal modeling for lip-reading[C] //Proceedings of the 30th British Machine Vision Conference. Durham: BMVA Press, 2019: Article No.276

[27] Feng D L, Yang S, Shan S G, *et al*. Learn an effective lip reading model without pains[OL]. [2023-07-26]. https://arxiv.org/abs/2011.07557

[28] Ma Jinlin, Liu Yuhao, Ma Ziping, *et al*. HSKDLR: lightweight lip reading method based on homogeneous self-knowledge distillation[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(11): 2689-2702(in Chinese)
(马金林, 刘宇灏, 马自萍, 等. HSKDLR: 同类自知识蒸馏的轻量化唇语识别方法[J]. 计算机科学与探索, 2023, 17(11): 2689-2702)

[29] Ma Jinlin, Liu Yuhao, Ma Ziping, *et al*. Lightweight lip reading method based on decoupling homogeneous self-knowledge distillation[J]. Journal of Beijing University of Aeronautics and Astronautics, 2024, 50(12): 3709-3719(in Chinese)
(马金林, 刘宇灏, 马自萍, 等. 解耦同类自知识蒸馏的轻量化唇语识别方法[J]. 北京航空航天大学学报, 2024, 50(12): 3709-3719)

[30] Ai X, Fang B. Cross-modal language modeling in multi-motion-informed context for lip reading[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2220-2232

[31] Ma Jinlin, Lyu Xin, Ma Ziping, *et al*. Micro-motion excitation and time perception for lip reading[J]. Acta Electronica Sinica, 2024, 52(11): 3657-3668(in Chinese)
(马金林, 吕鑫, 马自萍, 等. 微运动激励与时间感知的唇语识别方法[J]. 电子学报, 2024, 52(11): 3657-3668)

[32] Xu K, Li D W, Cassimatis N, *et al*. LCANet: end-to-end lipreading with cascaded attention-CTC[C] //Proceedings of the 13th IEEE International Conference on Automatic Face & Gesture Recognition. Los Alamitos: IEEE Computer Society Press, 2018: 548-555

[33] Margam D K, Aralikatti R, Sharma T, *et al*. LipReading with 3D-2D-CNN BLSTM-HMM and word-CTC models[OL]. [2023-07-26]. https://arxiv.org/abs/1906.12170

[34] Rastogi A, Agarwal R, Gupta V, *et al*. LRNeuNet: an attention based deep architecture for lipreading from multitudinous sized videos[C] //Proceedings of the International Conference on Computing, Power and Communication Technologies. Los Alamitos: IEEE Computer Society Press, 2019: 1001-1007

[35] Liu J L, Ren Y, Zhao Z, *et al*. FastLR: non-autoregressive lipreading model with integrate-and-fire[C] //Proceedings of the 28th ACM International Conference on Multimedia. New York: ACM Press, 2020: 4328-4336

[36] Kim M, Kim H, Ro Y M. Speaker-adaptive lip reading with user-dependent padding[C] //Proceedings of the 17th European Conference on Computer Vision. Heidelberg: Springer, 2022: 576-593

[37] Jeevakumari S A A, Dey K. LipSyncNet: a novel deep learning approach for visual speech recognition in audio-challenged situations[J]. IEEE Access, 2024, 12: 110891-110904