LipFormer：基于视觉地标 Transformer 的未见说话人唇读学习

薛峰、李宇、刘德音、谢银岑、吴林，*IEEE* 高级会员，Richang Hong*，*IEEE* 高级会员

arXiv:2302.02141v1 [cs.CV] 4 Feb 2023

摘要—唇读是指理解视频中说话人的语音，并进一步将其转化为自然语言。最先进的唇读方法在处理重叠说话人时表现出色，即说话人同时出现在训练集和推理集中。然而，由于训练库中的说话人数量有限，且不同说话人的嘴唇形状/颜色会造成明显的视觉变化，将这些方法泛化到未见说话人时，性能会急剧下降。因此，仅依赖嘴唇的可见变化容易导致模型过拟合。为解决这一问题，我们提出结合视觉信息与地标的多模态特征，这些特征能够描述唇部运动，而不受说话人身份的影响。随后，我们开发了一种基于视觉地标 Transformer 的句子级唇读框架，即*LipFormer*。具体而言，LipFormer 由唇部运动流、面部地标流和跨模态融合模块组成。两个流的嵌入由自注意力机制生成，并输入交叉注意力模块，以实现视觉信息与地标之间的对齐。最后，级联序列到序列模型可对生成的融合特征进行解码，从而输出文本。实验表明，我们的方法能够有效增强模型对未见说话人的泛化能力。

索引词—**唇读、地标、Transformer、唇部运动。**

I. 引言

唇读 IPREADING 是指根据视频片段推断说话人的语音，视频片段可以带有或不带有音频信号 [1]–[3]。唇读提供了一种有效的文本推断方式，可作为语音识别的替代方案；在某些干扰情境下，例如在野外遇到未知说话人时，语音识别可能难以实现。此外，唇读在无声电影处理和无声对话等现实应用中展现出巨大价值 [4]–[6]。

受益于深度学习，唇读也取得了显著进展，甚至展现出超越经验丰富的领域专家的趋势。早期工作仅在词级别上进行唇读 [7], [8]。然而，这类唇读方法每次只能对应一个词。与词级唇读相比，句子级唇读 [9]–[13] 通过依据上下文先验预测文本，在句子预测方面更加准确。例如，Assael 等人 [14] 提出了 LipNet，它结合了 VGG [15]，长短期记忆网络 [16],和连接时序分类 [17]，在 GRID 数据集 [18]上达到了 95.2 的准确率。在 [19], 中，作者开发了一种基于属性学习和对比学习的方法，大幅提升了唇读性能。然而，目前大多数唇读模型仅在公开数据集上进行训练和测试，而这些数据集的训练样本规模和说话人数都有限。此外，由于这些方法主要是在存在重叠说话人的情况下开发的（即测试说话人曾出现在训练集中），因此它们在未见说话人上的性能提升有限。我们假设，这些方法利用视觉线索描述唇部运动；然而，如果模型仅使用唇部视觉运动进行训练，就会因特定说话人的唇形/颜色和发音习惯所造成的视觉变化而过拟合。因此，这会阻碍模型的泛化能力。例如，如图 1 所示，模型若过拟合于视觉变化（如唇形），那么即使两位说话人说的是同一个词，也会将其转译为不同的文本。因此，单纯使用唇部运动的唇读方法可能会降低翻译准确率，尤其是在未见说话人的情况下。在实际应用中，唇读系统通常需要对训练库中可能未曾出现过的新面孔进行唇语到文本的预测。此外，学习一个能够泛化至未见说话人的模型，对下游应用至关重要。

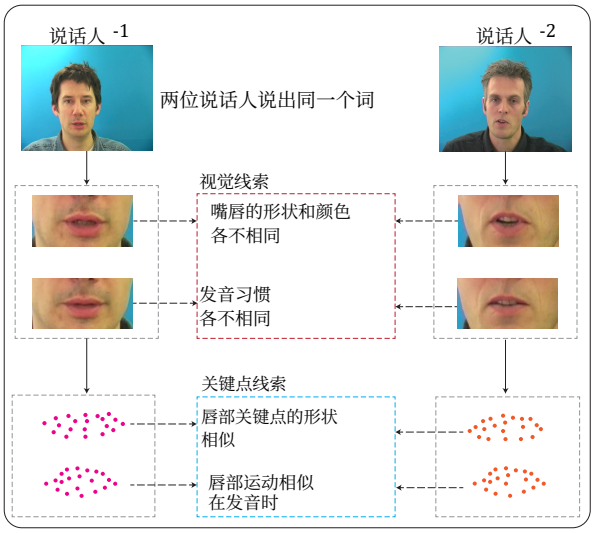


图 1. 唇部运动的视觉变化很容易导致唇读模型过拟合，因为模型会将运动与文本之间的虚假相关性联系起来。本文提出使用地标来校准跨模态关联。

鉴于可获取的样本数量和说话人

F.Xue、Y. Li、Y. Xie、L. Wu 和 R. Hong 均来自中国合肥工业大学。D. Liu 来自中国安徽大学。通讯作者：Richang Hong

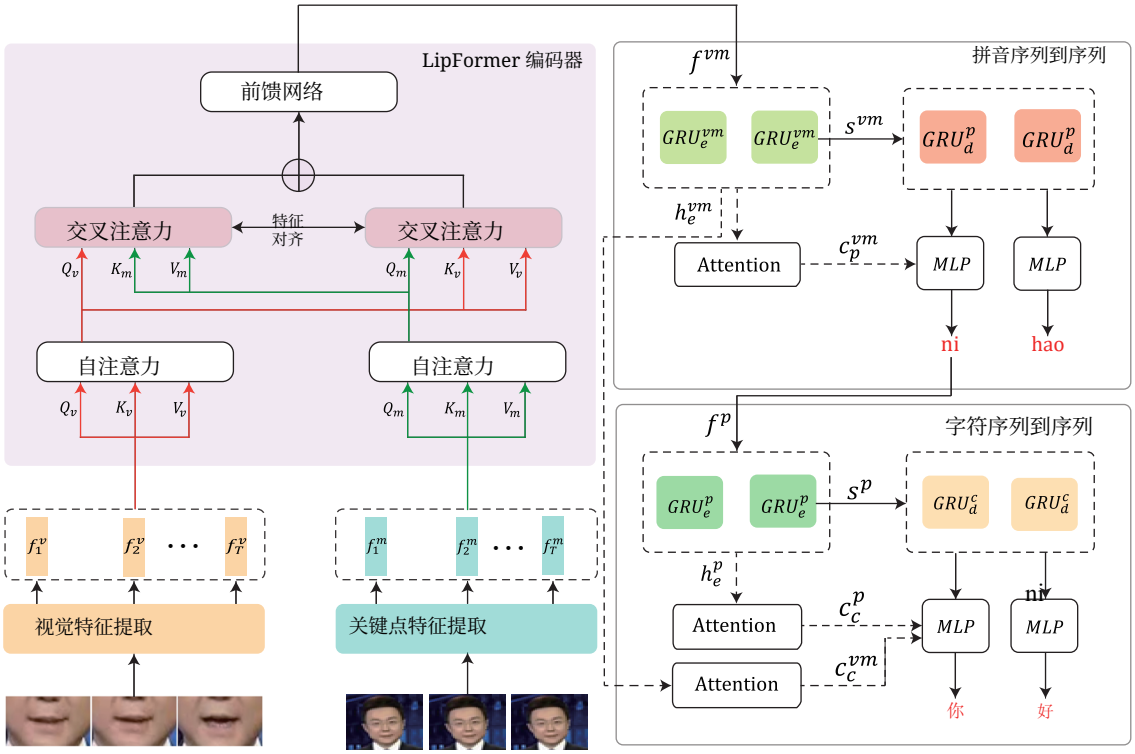


图2. 本文提出的 LipFormer 是一种基于视觉分支和关键点分支构建的端到端双流架构。视觉分支的输入是一系列唇部图像。关键点分支的输入是从说话人帧中提取的 340 维向量。两个流的嵌入被传入交叉注意力模块以实现对齐，从而实现跨模态融合。随后，级联序列到序列模型可对得到的融合特征进行解码，输出文本。符号说明见表I。

身份有限，我们旨在通过校正对齐校准运动与文本之间的关联，从而增强唇读模型，使其能够泛化到未见说话人。在本文中，我们探索使用地标作为校正偏移量，以实现唇部运动与唇读句子之间真实的潜在关联。在 [20],中，作者提出使用地标作为预训练深层表示的互补特征来学习物体结构表示，以识别视觉属性。在 [21],中，作者提出使用地标提取面部几何特征，几何特征和基于纹理的特征均可用于提高面部表情识别的准确率。这种方法还可以帮助模型泛化到训练集之外的新面孔。受这些方法启发，本文提出引入独立于嘴唇视觉外观的关键点特征，从而消除说话人之间的视觉变化。事实上，地标编码了说话人的面部和嘴唇的位置先验（即唇部和面部地标）。这些地标编码的运动轨迹能够有效描述唇部运动，而不受说话人影响。此外，学习到的关键点嵌入受视觉变化的影响较小。这表明，该方法有望在推理时提升唇读模型对未见说话人的泛化能力。

本文旨在提升唇读模型识别未见说话人的泛化能力。为实现这一目标，我们提出了一种基于视觉关键点变换器的句子级唇读框架，并将其命名为 LipFormer。具体而言，我们利用两种模态的特征描述唇部运动：视觉特征和地标。针对每位说话人，我们在帧级从唇部区域提取视觉特征，并通过自注意力机制使其具有判别性。然而，由于说话人唇部形状/颜色造成的视觉变化存在很强的偏差，仅使用视觉信息编码唇部运动很容易导致过拟合。为此，我们提出使用从面部和嘴唇提取的地标作为运动轨迹，以消除这类变化。随后，我们采用交叉注意力对齐并融合这些跨模态特征。Transformer 的交叉注意力能够有效学习视觉嵌入与地标嵌入之间的对应关系，从而改进用于跨模态融合的联合表示。最后，我们采用级联的序列到序列（seq2seq）模型解码融合特征并生成文本。

本文的主要贡献总结如下：

1) 我们提出了一种基于视觉关键点变换器的句子级唇读框架，该框架引入校正地标以最大限度地减少有偏的视觉变化，使模型能够泛化到未见说话人。

2）所提出的模型使用跨模态特征描述唇部运动，并采用交叉注意力实现视觉信息与地标之间的对齐，从而促进异构特征的融合，并进一步提升模型的

泛化能力。

3）在基准数据集上进行了大量实验，以证明所提出的方法在解读未见说话人方面的有效性，并取得了最先进的性能。

表I 符号说明及定义。

符号	定义
f_T^v, f_T^m	视觉特征，关键点特征
f^{vm}	视觉-关键点嵌入序列
f^p	拼音嵌入序列
GRU_e	下标 e 表示编码器
GRU_d	下标 d 表示解码器
GRU^{vm}	视觉-关键点编码器中的 GRU单元
GRU_e^p, GRU_d^p	拼音编码器和拼音解码器中的GRU单元
GRU_d^c	字符解码器中的GRU单元
h_e^{vm}, h_e^p	视觉-关键点编码器输出、拼音编码器输出
$c_p^{vm}, c_c^p, c_c^{vm}$	由注意力机制计算得到的上下文向量

II. 相关工作

本节将简要回顾唇读方法，这些方法可分为两类：传统方法和基于深度学习的唇读方法。

A. 传统唇读方法

1）基于像素的方法。该类方法假设唇部区域中的所有像素都包含与视觉相关的信息，并将唇部区域的像素值作为原始特征。通过不同方式对这些特征进行降维，以获得具有表达力的特征。例如，Potamianos 等人 [22] 提出了 HiLDA，该方法被广泛用作语音识别任务中的视觉特征提取器。Lucey 等人 [23] 在此基础上进一步考虑了局部特征，从图像块中提取局部特征，并将全局特征与局部特征融合，以进一步提高识别准确率。Tim 等人 [24] 对连续帧的 AAM特征进行归一化和拼接，并通过线性变换提取时空特征。2）基于形状的方法。该类方法根据唇部区域的形状（嘴唇、下巴等）提取特征。例如，Papcun 等人 [25] 使用发音特征（AFs）进行唇读，但由于这类特征过于简单，难以区分相似的单词，因此通常应用于小规模识别任务。Chan [26]将唇部的几何特征与PCA特征结合，作为视觉特征。Luettin 等人 [27] 将ASM模型应用于唇读，根据若干关键点的坐标生成特征。然而，基于形状的模型假设大部分与视觉相关的信息都位于由特征点表示的轮廓上，这不可避免地会导致信息丢失。

受益于深度学习，唇读也取得了显著进展。与传统方法相比，基于深度学习的方法具有强大的特征学习能力。深度学习方法无需复杂的手工特征提取流程，而且随着大规模数据的应用，其模型性能还能进一步提升。

B. 基于深度学习的唇读方法

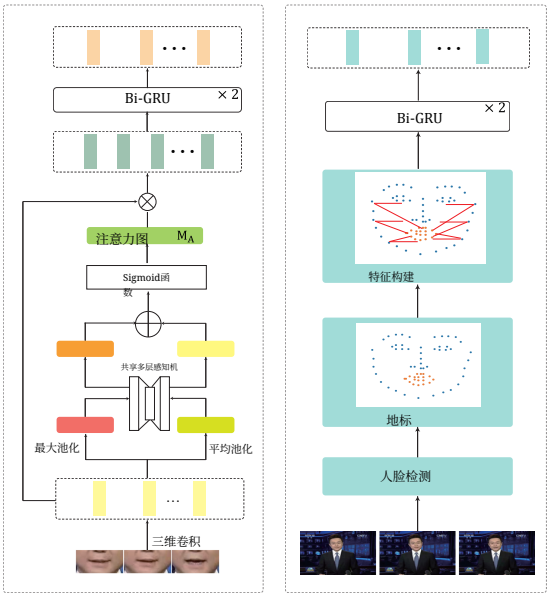
唇读可分为词级唇读和句级唇读。早期的研究仅在词级进行唇读，此类唇读视频仅对应一个词，应用范围较窄。例如，Chung [28] 设计了两种卷积神经网络结构，即早期融合和多塔结构，可根据一段唇部运动序列一次性完成词级翻译。Petridis 等人 [7] 提出了一种基于残差网络和双向门控循环单元的端到端音视模型，可同时直接从图像和音频中学习提取特征。Stafylakis 等人 [8]结合三维卷积神经网络和二维卷积神经网络提取视觉特征，并在LRW数据集上取得了更高的准确率。

本文重点研究句子级唇读。与词级唇读相比，句子级唇读的准确率更高，因为它会根据上下文先验来预测文本。LipNet [14]是首个端到端句子级唇读模型，由三维卷积神经网络、双向门控循环单元和连接时序分类组成。LipNet 在 theGRID 数据集上取得了 95.2% 的准确率。在 [29]–[31], 中，模型结构与 LipNet 相似。然而，CTC 损失具有条件独立性，即每个输出单元都独立地预测一个标签的概率。因此，CTC 损失会关注相邻帧的局部信息，不适用于预测需要借助上下文信息才能区分的标签。针对 CTC 损失带来的问题，Xu 等人 [12] 提出了 LCANet，该模型在三维卷积神经网络中堆叠了两层高速公路网络，从而显著提升了提取特征的质量。他们主要利用注意力机制来克服 CTC 中条件独立性假设的缺陷。后续工作 [19] 则通过在句子级唇读流程中引入属性学习和对比学习，提升了唇读模型的性能。

其他唇读方法基于序列到序列模型。最具代表性的模型是 WAS [32],，它使用 5 层二维卷积神经网络和长短期记忆网络提取视觉特征和听觉特征。这些特征被输入序列到序列模块以生成文本。Zhao [11]提出了 CSSMCM，该方法结合拼音、声调等因素，根据视觉信息帮助预测汉字。在 [1], 中，作者提出了一种知识蒸馏方法，使用经过预训练的语音识别模型作为教师模型，优化作为学生模型的唇读模型，从而提高唇读的准确率。

由于 Transformer 表现出色，Zhou 等人 [33]将 Transformer 应用于语音识别。Ma 等人[34] 将 Transformer 应用于唇读，并提出了 CTCH-LipNet。该方法首先使用三维卷积神经网络提取视觉特征，然后采用由两个 Transformer 组成的级联架构来预测拼音和汉字。Ma 等人 [35] 同时使用视频和音频作为输入，并通过 Transformer 将特征解码为文本。

然而，大多数现有唇读方法在实验评估中默认使用重叠说话人。这容易导致模型对训练集中的重叠说话人过拟合，而在未见说话人上的表现则较差。本文提出了一种基于视觉关键点变换器的句子级唇读



(a) (b)

图 3. 双流架构，包括视觉分支 (a) 和关键点分支 (b)。输入为唇部区域序列，我们使用结合通道注意力的三维卷积神经网络提取特征，然后通过双向门控循环单元编码时间顺序。关键点分支的输入是关键点嵌入，即相邻帧之间角度矩阵的差值。每帧中，角度矩阵由唇部的 20 个地标与面部轮廓的 17 个地标计算得出，并编码为一个 340 维向量。

框架，使模型能够泛化到未见说话人，从而解决未见说话人的唇读问题。

III. 方法

我们提出的 LipFormer 架构如图 2 所示，由四个模块组成：1）视觉流，用于提取唇部区域的视觉特征；2）标志点流，用于编码说话者在整个序列中的唇部/面部运动轨迹；3）跨模态融合，用于学习视觉信息与地标之间的对齐；4）级联序列到序列模型，用于将融合特征映射为文本。

A. 视觉嵌入

对于每个视频片段，我们首先使用DLib人脸检测器提取每一帧中的人脸图像，然后对每张人脸图像进行仿射变换，得到以嘴部为中心的裁剪区域，其中 160 × 80 像素为唇部区域。对于包含 T 帧的视频片段，我们可以得到一个唇部区域序列$\{T_i\}$，其中 I_i 是第 i步的帧（$i = 1, ……, T$）。我们首先通过对三维卷积 [36], 应用ReLU层和最大池化层，学习逐帧特征嵌入。训练期间，将丢弃正则化与三维卷积神经网络结合使用，以缓解饱和问题。

三维卷积神经网络提取的视觉特征包含大量无关信息，例如唇形、不同说话人的发音习惯等。这些信息会严重影响解码器生成文本的准确性。本文将三维卷积神经网络与通道注意力相结合。通道注意力机制能够学习各个通道的权重，并通过抑制无关特征来提升有效视觉特征的表现。每一帧获得的特征向量记为 f_i^v。

为了聚合空间网格，使用平均池化和最大池化从输入特征图中生成两种不同的空间上下文描述符，并将这两个输出输入共享的多层感知机网络，以生成通道注意力图。将注意力图与输入特征相乘，以聚焦于重要特征。视觉分支的结构如图3(a)所示。

受卷积核大小的限制，卷积神经网络只能提取短期时空视觉特征。我们采用双向门控循环单元 [37], [38] 来提取长期特征：

$$[s_i^v, o_i^v] = GRU(f_i^v, s_{i-1}^v), \tag{1}$$

其中，s_i^v、o_i^v 分别是门控循环单元的隐藏向量和输出向量。

B. 地标嵌入

仅依赖视觉外观会导致唇读性能不佳。主要原因是，唇部运动存在显著的视觉变化，这是由嘴唇形状和颜色的差异，或说话人特定的发音习惯造成的。为消除视觉变化，我们提出将地标作为另一种特征嵌入。关键点嵌入受唇部外观的影响较小，在训练过程中对未见说话人具有更好的泛化能力。图 3(b) 展示了 68 个面部地标的分布，其中嘴唇有 20 个地标，面部轮廓有 17 个地标，眼睛、眉毛和鼻子共有 31 个地标。关键点分支的结构如图 3(b) 所示。

更具体地说，我们利用唇部地标和面部轮廓地标来构建用于唇读的关键点嵌入。研究发现，面部轮廓位置的变化与唇部运动最为明显。面部地标之间的运动轨迹能够有效描述唇部运动。特征构建如下：1）首先，我们计算每一帧中唇部的20个地标与面部轮廓的17个地标之间的夹角，即余弦值，从而得到维度为 (1,340) 的角度矩阵；2）随后，我们计算相邻帧之间的角度矩阵差值，以表示地标的运动变化。3）对于T帧，得到维度为 (B,T,340) 的角度矩阵，作为关键点分支的输入。矩阵差值即为逐帧地标嵌入，记为f_i^m，对应第 i帧。给定所有 T帧的 $\{f_1^m, f_i^m, ...f_T^m\}$，我们使用双向门控循环单元提取长期特征，从而得到输出特征序列$\{s_1^m, s_i^m, ...s_T^m\}$。

C. 基于 *Transformer* 的跨模态融合

来自两个流的嵌入被输入 Transformer。我们使用 Transformer编码器实现跨模态融合。编码器由一个编码器层组成，该层包含三个部分：自注意力模块、交叉注意力模块和前馈网络。两个流的嵌入由自注意力生成，

然后输入交叉注意力模块，以实现视觉信息与地标之间的对齐。查询 Q、键 K 和值 V 这三个矩阵作为自注意力的输入，由输入序列 z 生成。自注意力模块提取全局信息，以建立唇部运动的长期全局依赖关系。交叉注意力将当前模态（例如视频）的 Q，以及从另一模态（例如地标）获得的 K/V 作为输入；对于每个视觉特征嵌入，交叉注意力会为每个地标特征嵌入分配不同的权重，通过权重匹配实现视觉-地标嵌入对齐，从而实现跨模态特征融合。注意力模块的输出为：

$$Attention = \textbf{softmax}\left(\frac{Q^i(K^i)^T}{\sqrt{d}}\right)\cdot V^i. \tag{2}$$

其中 d 是嵌入向量的长度。

编码器层采用前馈网络实现，包含两个带有 ReLU 非线性的全连接层：

$$FFN(x) = FC(ReLU(FC(x))). \tag{3}$$

D. 文本生成

我们将融合特征作为输入，送入级联序列到序列模型。该序列到序列模型采用编码器-解码器结构，其中编码器和解码器均为长短期记忆网络模型（有时也使用门控循环单元模型）。编码器-解码器模型可以预测任意序列之间的对应关系。在中文数据集中，拼音类别较少，因此更容易预测拼音。基于此，我们在预测汉字时选择拼音作为中间层，并使用级联序列到序列模型解码文本。

E. 拼音预测

我们将融合特征 f^{vm} 作为输入，送入拼音序列到序列模型，将 f^{vm} 解码为拼音。我们将编码器和解码器分别称为视觉-关键点编码器和拼音解码器，其中编码器处理视觉-关键点特征序列，解码器预测拼音序列。编码器中的门控循环单元通过输入 f^{vm} 计算隐藏层向量：

$$(h_e^{vm})_i = GRU_e^{vm}((h_e^{vm})_{i-1}, e_i^{vm}). \tag{4}$$

解码器根据前一时间步的预测结果 p_i 和前一时间步的隐藏状态（h_d^p）$_{i-1}$，逐步计算隐藏层向量（h_d^p）$_i$：

$$(h_d^p)_i = GRU_d^p((h_d^p)_{i-1}, E(p_i)), \tag{5}$$

其中，嵌入函数 $E(\cdot)$ 将 p_i 映射到其向量空间。为充分利用输入向量中包含的信息，我们使用注意力模块计算编码器中所有隐藏层向量的权重，以生成上下文向量，帮助解码器预测拼音。权重计算如下：

$$att = \textbf{softmax}(FC(tanh(FC(Concat((h_d^p)_i), h_e^p)))). \tag{6}$$



图 4. CMLR 数据集中的静态图像。

最后，将 GRU 输出与上下文向量拼接，计算当前拼音的概率分布，如下所示：

$$P(p_i) = \textbf{softmax}(FC((h_d^p)_i, h_e^p \cdot att. \tag{7}$$

F. 汉字预测

同样地，以拼音序列作为输入，将其送入字符序列到序列模型，将拼音转换为汉字。其计算过程与预测拼音序列类似。首先，将拼音序列输入拼音编码器，拼音编码器中的门控循环单元计算隐藏层向量如下：

$$(h_e^p)_i = GRU_e^p((h_e^p)_{i-1}, e_i^p). \tag{8}$$

此外，为了同时利用视觉地标和拼音信息，在预测汉字的过程中采用了双重注意力机制。字符解码器根据视觉-关键点编码器的输出和拼音编码器的输出计算上下文向量，然后预测汉字：

$$(h_d^c)_i = GRU_d^c((h_d^c)_{i-1}, E(c_i)). \tag{9}$$

$$c_i^{vm} = h_e^{vm} \cdot att((h_d^c)_i, h_e^{vm})). \tag{10}$$

$$c_i^p = h_e^p \cdot att((h_d^c)_i, h_e^p)). \tag{11}$$

$$P(c_i) = \textbf{softmax}(FC((h_d^c)_i, c_i^{vm}, c_i^p). \tag{12}$$

G. 损失函数

为提高预测准确率，模型首先预测拼音，然后将拼音转换为汉字。我们联合优化这两个过程的损失函数。损失函数定义如下：

$$L = L_p + L_c, \tag{13}$$

where $L_p = -\sum_{n=1}^{N} logP(p_n|x, p_1, p_2, ...p_{n-1})$, $L_c = -\sum_{n=1}^{N} logP(c_n|x, c_1, c_2, ...c_{n-1})$

IV. 实验

本节通过在两个基准数据集上开展大量实验，验证所提出的方法：CMLR[11]和 GRID [18]。

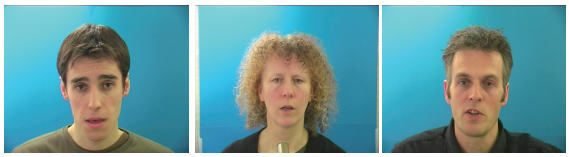


图 5. theGRID 数据集中的静态图像。



图 6. 以嘴部为中心的裁剪图示例。

A. 数据集

中文普通话唇读（CMLR）数据集 [11]是规模最大的汉语普通话句子唇读数据集（示例图像见图 4）。整个数据集包含由 11 位说话人录制的序列。我们将数据集划分为训练集和测试集，分别包含 9 位和 2 位说话人。请注意，训练集和测试集中的说话人不重叠。针对说话人重叠的实验，我们按照 [11]划分训练集、验证集和测试集。CMLR 划分协议见表 II。

表 II CMLR 数据集的统计信息。

	Set	说话人	句子
未见	训练	9	81094
	Test	2	20978
重叠	训练	11	71452
	Test	11	20418

GRID 数据集 [18] 录制了 33 位说话人的数据，是唇读领域广泛使用的数据集（示例图像见图 5）。每个句子由一连串的动词 +颜色 + 介词 + 字母 + 数字 + 副词组成，例如“蓝色垃圾桶，蓝色，在 f 五，再来一次”。为了将数据集划分为未见说话人和重叠说话人，我们遵循了 [14] 中建议的设置。划分方案见表 III。

表 III GRID 数据集的统计信息。

	Set	说话人	句子
未见	训练	29	28837
	Test	4	3986
重叠	训练	33	24408
	Test	33	8415

B. 评估指标

为衡量所提方法和基线方法的性能，我们采用自动语音识别中广泛使用的评估指标：词错误率（WER）和字符错误率（CER）。WER/CER 定义为将预测标签转换为真实值所需的最少词语/字符操作次数（包括替换、删除和插入操作），再除以真实值中的词语/字符数。计算公式如下：

$$WER/CER = 100 \cdot \frac{S + D + I}{N}, \tag{14}$$

其中，S 表示替换，D 表示删除，I 表示插入，N 表示真实值中的词语数。请注意，WER/CER 越小，预测准确率越高。此外，我们在中文数据集 CMLR 上使用的评估指标仅为 CER，其中 CER 表示字符错误率。CER 的计算方式与 WER 相同，即将每个汉字视为一个英文单词。

C. 实现细节

对于每个视频片段，我们首先使用 DLib人脸检测器 [39]提取每一帧中的人脸图像。人脸检测器生成的地标坐标用作关键点分支的输入。对每张人脸图像应用仿射变换，得到 160 × 80像素的以嘴部为中心的裁剪图，作为视觉分支的输入。（见图 6）。本实验使用 Adam优化器优化参数，初始学习率为 0.0003。当训练误差连续 4 个训练轮次未改善时，学习率降低 50%。

D. 竞品

我们将所提出的方法与几种最先进方法进行比较：Lip Net [14],CSSMCM [11], CALLip [19], LCSNet [40]和 WAS [32]。

LipNet：LipNet 是首个端到端句子级唇读模型，在 GRID 数据集上达到 95.2 的准确率。

CSSMCM：该模型专为中文唇读设计，结合拼音和声调等因素，帮助预测汉字。

CALLip：CALLip 将属性学习和对比学习引入句子级唇读流程，以提升模型性能；该流程使用属性学习模块提取说话人身份特征，并消除跨说话人差异。

LCSNet：该模型通过通道注意力模块和选择性特征融合模块，提取与唇部运动更相关的特征，从而提高识别准确率。

WAS：该基线模型使用视频信息，通过序列到序列模型预测句子，并在LRS数据集上取得了先进的性能。

E. 与竞品的比较

为了验证所提出方法的有效性，我们首先将其与 CMLR 和

表IV CMLR 上与最先进方法的性能比较。-：暂无结果。最佳结果以粗体显示。

方法	未见	重叠
LipNet	52.18	33.41
WAS	-	38.93
CSSMCM	50.08	32.48
CALLip	-	31.18
LCSNet	46.98	30.03
LipFormer	**43.18**	**27.79**

表 V GRID 上与当前最先进方法的性能比较。最佳结果以粗体显示。

方法	未见	重叠
LipNet	17.5	4.8
WAS	14.6	3.0
CALLip	-	2.48
LCSNet	11.6	2.3
LipFormer	**9.64**	**1.45**

GRID 数据集上的 SOTA 竞品进行比较。我们通过实验证实，CTC 损失可能导致 CMLR 数据集训练过程中不收敛。因此，我们用级联 Seq2Seq 模块替换了连接时序分类。我们报告了两个数据集中未见说话人和重叠说话人的实验结果。实验结果分别如表 IV 和表 V 所示。

表IV显示，与最先进的方法相比，LipFormer 在未见说话人上的表现有显著提升。与 LipNet 相比，LipFormer 使用地标共同描述唇部运动。对比结果表明，采用多模态特征有助于提升性能。与 CSSMCM 相比，LipFormer 的词错误率进一步降低了 6.9%。原因之一是 LipFormer 利用多模态特征学习到了更加一致的唇部运动特征，因此模型能够很好地泛化到未见说话人。对于重叠说话人，LipFormer 在 CMLR 数据集上优于最先进的方法，字符错误率达到 27.79%。这表明 LipFormer 非常适用于未见说话人和重叠说话人。

表V展示了不同方法在GRID数据集上的比较结果。与LipNet相比，LipFormer使未见说话人的词错误率降低了7.86%，使重叠说话人的词错误率降低了3.35%。这证实了地标有助于解决泛化问题。还可以观察到，LipFormer在未见说话人和重叠说话人上的表现均优于其他方法，尽管这些说话人来自不同族群。

F. 消融研究

为验证关键点分支和Transformer模块能否有效提高未见说话人的唇读准确率，我们将LipFormer框架拆分，并在CMLR和GRID数据集上进行测试。实验结果分别见表VI和表VII。图7和图8展示了不同模型变体在两个数据集上的词错误率曲线。

表VI CMLR上LipFormer及其变体的性能。前端结构的

LIPFORMER 是纯视觉+地标+TRANSFORMER。最佳结果以粗体显示。

#	方法	未见	重叠
1	纯视觉	48.14	29.15
2	纯视觉+地标	43.48	28.0
3	LipFormer	**43.18**	**27.79**

表VII LipFormer及其变体在GRID上的性能。最佳结果以粗体显示。

#	方法	未见	重叠
1	纯视觉	12.35	2.82
2	纯视觉+地标	10.24	2.2
3	LipFormer	**9.64**	**1.45**

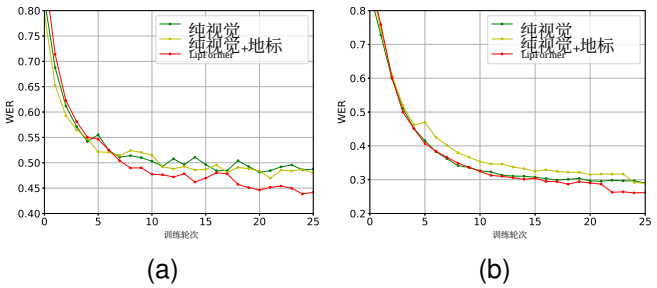


图7. CMLR 数据集上模型变体的性能。(a) 未见。(b) 重叠。

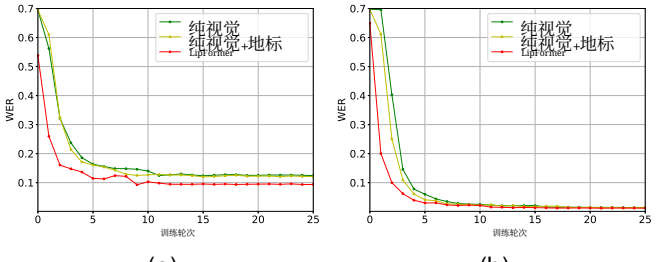


图8. GRID 数据集上模型变体的性能。(a) 未见。(b) 重叠。

我们方法的变体，即 #1 和 #2，通过融合视觉分支和关键点分支，优于采用纯视觉分支的模型。具体而言，与纯视觉模型相比，对于未见说话人和重叠说话人，方法 # 2 在 CMLR 数据集上的词错误率分别降低了 4.66% 和 1.15%，在 GRID 数据集上分别降低了 2.11% 和 0.62%。这表明，关键点能够规范不同说话人的唇形，消除无关的视觉变化，并增强模型对未见说话人的泛化能力。将 LipFormer 与方法 #2 进行比较，实验结果表明，Transformer 模块学习视觉-地标嵌入之间的对应关系，从而实现跨模态异构

表VIII CMLR 数据集上不同嵌入维度的模型性能。

#	方法	未见	重叠
1	256	43.61	28.34
2	512	**43.18**	**27.79**
3	1024	43.47	28.12

特征的融合，进一步提升模型性能。

G. 对超参数的敏感性

本节在 CMLR 数据集上进行实验，以研究 LipFormer 中超参数对模型性能的影响。特征提取器将每个时间步编码为一个特征嵌入。为研究关键点分支中特征嵌入的大小如何影响模型的学习性能，我们通过控制关键点分支中门控循环单元（GRU）的输出通道数来调整嵌入维度。我们设置了三种不同的维度：256、512 和 1024。

表VIII总结了不同嵌入维度下各模型的性能。总体而言，随着嵌入维度增加，模型性能会有所提升。然而，嵌入维度过大可能导致模型过拟合，反而使模型性能下降。对于 CMLR 数据集，无论是未见说话人还是重叠说话人，GRU 第一层输出的最佳通道数均为 512。

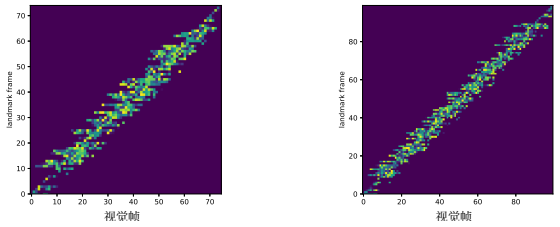
H. 案例研究

为定性分析所提出模型的性能，本节评估部分预测结果。表 IX 展示了各变体模型在 CMLR 和 GRID 数据集上生成的一些句子，其中红色标出的汉字为错误内容。

从表 IX 可以看出，纯视觉模型生成的句子与 CMLR 数据集上的真实值存在一些差异，例如，将“人口”（population）预测为“中外”（Chinese-foreign），将“一些”（some）预测为“议程”（agenda）等。这表明，由于唇形等视觉变化，即使人们说的是同一个词，模型也会识别出不同的文本。纯视觉+地标模型生成的句子更接近真实值，这说明地标能够提高模型识别不同说话人说出同一句话时的准确率。LipFormer 生成的句子均正确，这表明 Transformer 能够促进跨模态信息的融合。在 GRID 数据集上，有些字母容易被预测错误，例如将“f”预测为“s”。这些错误主要是由发音相近的字母造成的。

I. 对齐可视化

图9a和图9b分别展示了交叉注意力模块在 CMLR 数据集上对 75 帧和 100 帧处视觉信息与地标之间对齐情况的可视化。图中每一行代表一种视觉模态，每一列代表一种关键点模态。图中的高亮区域表示特征融合过程中视觉-关键点特征嵌入之间的对齐程度。



(a) (b)

图 9. CMLR 数据集上视觉-地标嵌入对齐示意图。纵轴表示视觉嵌入，横轴表示关键点嵌入。

图 9 显示，在特征融合过程中，跨模态注意力机制能够学习两种模态特征嵌入之间的对应关系，实现视觉-地标嵌入之间的对齐，其对角线趋势十分明显。在发音相同的词语时，注意力会集中于不同模态中对应的帧。

V. 结论

本文提出了一种用于句子级唇读的跨模态 Transformer 框架，通过将地标用作运动轨迹来校准视觉变化，从而泛化到未见说话人。该模型通过交叉注意力所提出的跨模态融合，提升了异构特征的对齐效果。大量实验结果表明，我们的框架能够有效泛化到未见说话人。未来，我们将进一步探讨这项研究在更具挑战性的数据集（例如侧视说话人）上的相关性。

参考文献

[1] Y. Zhao, R. Xu, X. Wang, P. Hou, H. Tang, and M. Song, “听见唇语：通过蒸馏语音识别器改进唇读”，2019年。[2] X. Zhang, H. Gong, X. Dai, F. Yang, N. Liu, and M. Liu, “利用面部特征理解象形文字：端到端中文句子级唇读”，载于人工智能促进协会年会论文集，第33卷，第01期，2019年，第9211–9218页。[3] S. Hilder, R. W. Harvey, and B.-J. Theobald, “人类与机器唇读的比较”，载于*AVSP*，2009年，第86–89页。[4] J. O. Kim, W. Lee, J. Hwang, K. S. Baik, and C. H. Chung, “基于多分辨率架构的安全系统唇纹识别”，*Future Generation Computer Systems*，第20卷，第2期，第295–301页，2004年。[5] O. K. Jin, W. Lee, J. Hwang, K. S. Baik, and C. H. Chung, “基于多分辨率架构的安全系统唇纹识别”，*Future Generation Computer Systems*，第20卷，第2期，第295–301页，2004年。[6] B. Xu, C. Lu, Y. Guo, and J. Wang, “判别式多模态语音识别”，2020年。[7] S. Petridis, T. Stafylakis, P. Ma, F. Cai, G. Tzimiropoulos, and M. Pantic, “端到端视听语音识别”，载于*2018 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2018年，第6548–6552页。[8] T. Stafylakis and G. Tzimiropoulos, “将残差网络与长短期记忆网络结合用于唇读”，*arXiv*预印本 *arXiv:1703.04105*，2017年。

表 IX CMLR 和 GRID 上不同模型的预测结果。

方法	CMLR		GRID	
	未见	重叠	未见	重叠
真实值	人口较少民族发展规划	一些新做法取代了老传统	请在 f 三处放蓝色箱子	尽快在 d 二处放蓝色箱子
纯视觉	中外较少数民族发展模式	议程改进取代了措施	蓝色垃圾桶在三点，请	在蓝色垃圾桶旁尽快
纯视觉+地标	人口较少数民族人口普查	一些新做法取代了老烧	蓝色垃圾桶在l 三号，请	蓝色垃圾桶在 d 二号再来一次
LipFormer	人口较少民族发展规划	一些新做法取代了老传统	请在 f 三处放蓝色箱子	尽快在 d 二处放蓝色箱子

[9] T. Zhang, L. He, X. Li, and G. Feng, “利用时间卷积网络实现高效的端到端句子级唇读”，《应用科学》，第11卷，第15期，第6975页，2021年。[10] X. Zhao, S. Yang, S. Shan, and X. Chen, “通过互信息最大化实现有效唇读”，载于*2020*年第*15*届*IEEE*自动人脸与手势识别国际会议（*FG 2020*）。IEEE，2020年，第420–427页。[11] Y. Zhao, R. Xu, and M. Song, “用于汉语普通话唇读的级联序列到序列模型”，载于*ACM*亚洲多媒体会议论文集，2019年，第1–6页。[12] K. Xu, D. Li, N. Cassimatis, and X. Wang, “LCANet：采用级联注意力CTC的端到端唇读”，载于*2018*年第*13*届*IEEE*自动人脸与手势识别国际会议（*FG 2018*）。IEEE，2018年，第548–555页。[13] J. Liu, Y. Ren, Z. Zhao, C. Zhang, B. Huai, and J. Yuan, “FastLR：采用积分发放的非自回归唇读模型”，载于第*28*届*ACM*国际多媒体会议论文集，2020年，第4328–4336页。[14] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, “LipNet：端到端句子级唇读”，*arXiv*预印本 *arXiv:1611 01599*，2016。。[15] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “细节中的魔鬼归来：深入探究卷积网络”，*arXiv*电子预印本，2014年。[16] J. Chung, C. Gulcehre, K. H. Cho, and Y. Bengio, “门控循环神经网络在序列建模中的实证评估”，*arXiv*预印本，2014年。[17] A. Auvolat and T. Mesnard, “连接时序分类：使用循环神经网络为未分段序列标注”，2006年。[18] M. Cooke, J. Barker, S. Cunningham, and X. Shao, “用于语音感知和自动语音识别的视听语料库”，《美国声学学会杂志》，第120卷，第5期，第2421–2424页，2006年。[19] Y. Huang, X. Liang, and C. Fang, “CALLip：利用对比学习和属性学习进行唇读”，载于第*29*届*ACM*国际多媒体会议论文集，2021年，第2492–2500页。[20] Y. Zhang, Y. Guo, Y. Jin, Y. Luo, Z. He, and H. Lee, “无监督发现物体地标作为结构表示”，载于*IEEE*计算机视觉与模式识别会议论文集，2018年，第2694–2703页。[21] M. A. Haghpanah, E. Saeedizade, M. T. Masouleh, and A. Kalhor, “利用面部地标和神经网络进行实时面部表情识别”，载于*2022*年机器视觉与图像处理国际会议（*MVIP*）。IEEE，2022年，第1–7页。[22] V. Estellers, M. Gurban, and J. P. Thiran, “关于视听语音识别中的动态流加权”，*IEEE*音频、语音与语言处理汇刊，第20卷，第4期，第1145–1157页，2012年。[23] P. J. Lucey, G. Potamianos, and S. Sridharan, “多视角视觉语音的基于图像块分析”，载于*2008*年听觉-视觉语音处理国际会议论文集，2008年。[24] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, “非言语交际感知回归中的文化因素”，载于*2011*年*IEEE*国际计算机视觉会议研讨会（*ICCV*研讨会）。IEEE，2011年，第1242–1249页。[25] G. Papcun, J. Hochberg, T. R. Thomas, F. Laroche, and S. Levy, “利用在X射线微束数据上训练的神经网络，从声学信息推断发音动作并识别手势”，《美国声学学会杂志》，第92卷，第2期第1部分，第688–700页，1992年。[26] M. T. Chan, “融合几何视觉特征与基于外观的视觉特征的基于隐马尔可夫模型的视听语音识别”，载于*IEEE*第四届多媒体信号处理研讨会，2001年。[27] J. Luettin and N. A. Thacker, “利用概率模型进行读唇”，《计算机视觉与图像理解》，第65卷，第2期，第163–178页，1997年。[28] J. S. Chung 和 A. Zisserman，《野外唇读》，载于计算机视觉——*ACCV 2016*，2017年，第 87–103 页。[29] I. Fung 和 B. Mak，《使用 Maxout 卷积神经网络和长短期记忆网络实现端到端低资源唇读》，2018年，第 2511–2515 页。[30] M. Wand 和 J. Schmidhuber，《通过域对抗训练改进说话人无关的唇读》，2017年。[31] M. Wand、J. Schmidhuber 和 N. T. Vu，《端到端视听融合研究》，2018年。[32] J. Son Chung、A. Senior、O. Vinyals 和 A. Zisserman，《野外唇读句子》，载于*IEEE*计算机视觉与模式识别会议论文集，2017年，第 6447–6456 页。[33] S. Zhou、L. Dong、S. Xu 和 B. Xu，《基于音节的 Transformer 普通话语音识别序列到序列模型》，*arXiv* 预印本 *arXiv:1804.10752*，2018年。[34] S. Ma、S. Wang 和 X. Lin，《面向句子级普通话唇读的基于 Transformer 的模型》，载于*2020* 年 *IEEE* 第五届网络空间数据科学国际会议（*DSC*），2020年。[35] P. Ma、S. Petridis 和 M. Pantic，《野外多种语言的视觉语音识别》，2022年。[36] L. Wu、Y. Wang、L. Shao 和 M. Wang，《3D PersonVLAD：学习用于基于视频的行人重识别的深层全局表征》，*IEEE* 神经网络与学习系统汇刊，第 30 卷，第 11 期，第 3347–3359 页，2019年。[37] L. Wu、Y. Wang、J. Gao、M. Wang、Z.-J. Zha 和 D. Tao，《用于行人重识别相对表征学习的深度协同注意力比较器》，*IEEE* 神经网络与学习系统汇刊，第 32 卷，第 2 期，第 722–735 页，2021年。[38] D. Chen、M. Wang、H. Chen、L. Wu、J. Qin 和 W. Peng，《基于异构图嵌入的跨模态检索》，载于第 *30* 届 *ACM* 多媒体国际会议论文集，2022年，第3291–3300 页。[39] A. Amodio、M. Ermidoro、D. Maggi、S. Formentin 和 S. M. Savaresi，《基于瞳孔对光反射的驾驶员功能受损自动检测》，*IEEE* 智能交通系统汇刊，第 20 卷，第 8 期，第 3038–3048 页，2018年。[40] F. xue、T. Yang、K. Liu、Z. Hong、M. Cao、D. Guo 和 R. Hong，《LCSNet：通过通道感知特征选择实现端到端唇读》，*ACM* 多媒体计算、通信与应用汇刊（*TOMM*），2022年。