基于Transformer和多模态对齐的非自回归手语翻译技术研究

邵舒羽*① 杜 垚② 范晓丽③
①(北京物资学院物流学院 北京 101149)
②(北京航空航天大学自动化科学与电气工程学院 北京 100191)
③(空军特色医学中心 北京 100142)

摘 要：为了解决多模态数据的对齐及手语翻译速度较慢的问题，该文提出一个基于自注意力机制模型Transformer的非自回归手语翻译模型(Trans-SLT-NA)，同时引入了对比学习损失函数进行多模态数据的对齐，通过学习输入序列(手语视频)和目标序列(文本)的上下文信息和交互信息，实现一次性地将手语翻译为自然语言。该文所提模型在公开数据集PHOENIX-2014T(德语)、CSL(中文)和How2Sign(英文)上进行实验评估，结果表明该文方法相比于自回归模型翻译速度提升11.6～17.6倍，同时在双语评估辅助指标(BLEU-4)、自动摘要评估指标(ROUGE)指标上也接近自回归模型。

关键词：手语翻译；自注意力机制；非自回归翻译；深度学习；多模态数据对齐
中图分类号：TN108.4; TP391 **文献标识码**：A **文章编号**：1009-5896(2024)07-2932-10
DOI: 10.11999/JEIT230801

Non-Autoregressive Sign Language Translation Technology Based on Transformer and Multimodal Alignment

SHAO Shuyu① DU Yao② FAN Xiaoli③
①(*School of Logistics*, *Beijing Wuzi University*, *Beijing* 101149, *China*)
②(*School of Automation Science and Electrical Engineering*, *Beihang University*, *Beijing* 100191, *China*)
③(*Air force medical center*, *PLA*, *Beijing* 101142, *China*)

Abstract: To address the challenge of aligning multimodal data and improving the slow translation speed in sign language translation, a Transformer Sign Language Translation Non-Autoregression (Trans-SLT-NA) is proposed in this paper, which utilizes a self-attention mechanism. Additionally, it incorporates a contrastive learning loss function to align the multimodal data. By capturing the contextual and interaction information between the input sequence (sign language videos) and the target sequence (text), the proposed model is able to perform sign language translation to natural language in s single step. The effectiveness of the proposed model is evaluated on publicly available datasets, including PHOENIX-2014-T (German), CSL (Chinese) and How2Sign (English). Results demonstrate that the proposed method achieves a significant improvement in translation speed, with a speed boost ranging from 11.6 to 17.6 times compared to autoregressive models, while maintaining comparable performance in terms of BiLingual Evaluation Understudy (BLEU-4) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metrics.

Key words: Sign language translation; Self-attention mechanism; Non-autoregressive translation; Deep learning; Alignment of multimodal data

收稿日期：2023-08-01；改回日期：2023-12-27；网络出版：2024-01-08
*通信作者： 邵舒羽 shaoshuyu@bwu.edu.cn
基金项目：国家自然科学基金(8210072143)，北京市教委科技计划青年项目(KM202210037001)
Foundation Items: The National Natural Science Foundation of China (8210072143), R&D Program of Beijing Municipal Education Commission (KM202210037001)

1　引言

手语(sign language)是听障人士用来与他人交流的主要方式，他们通过手语来感知世界并学习技能。然而不同于自然语言，手语的语义主要包含在面部表情、手势动作、眼神和唇型等[1]，非手语使用人群往往无法完全准确地理解手语信息，因此与手语使用者之间的交流沟通存在障碍[2]。由于沟通不畅，学习手语需要花费大量的精力和时间成本，建立相关的算法使计算机进行手语翻译可以方便人们之间的交流，提高手语翻译任务技术也引起了研究人员广泛的兴趣。

手语翻译目标为将一连串连续的手语视频经过模型和算法的处理后，将其转化为对应的自然语言文本[3]。手语翻译在当前的研究中也被认为是一种序列到序列(seq-to-seq)的任务[4]，相比于机器翻译任务，手语翻译的输入和输出分别是视频和文本，属于多模态的任务。而视频与文本各自语义信息的表达方式不同，因此要实现视觉语义信息跨模态转换为自然语言语义表达，需要通过能够同时处理图像和文本数据的深度学习图文大模型来实现。

目前对于手语翻译任务的研究主要是将视觉空间信息提取和时间序列表达学习两部分相结合[5,6]。与视频动作检测任务不同，手语翻译任务没有明确的动作边界，不同的手语标注文本(gloss)之间存在空白填充(blank)或重叠交叉，同时还需要进行序列表征学习来使模型充分理解视频信息，目前的研究主要基于二维卷积神经网络(Two-Dimensional-Convolutional Neural Network, 2D-CNN)[7]、长短期记忆网络(Long Short-Term Memory, LSTM)[8]等模型进行序列表征学习和对上下文信息的理解，从而将视频转化为对应的文本序列。2017年谷歌提出基于自注意力机制的模型Transformer在机器翻译任务中取得了当时最好的效果[9]，表明了自注意力机制处理序列数据的强大能力。在手语翻译中，研究人员也将该模型引入同时实现了高质量的手语识别和手语翻译[10]。

学者们相继提出了基于Transformer的联合手语转换模型手语转换器(Sign Language Transformer, SLT)[5]、深度Transformer模型[11]、多模态学习模型[12]等手语翻译图文大模型，对于手语翻译的研究取得较大的进展，但仍存在较为显著的问题。首先是手语翻译效率问题，以Transformer为基础的模型虽然可以实现高质量的翻译，但该模型是编码器-解码器结构，其逐词生成模式对于较长序列的翻译速度缓慢[13]，并且计算开销较大，不利于在边缘设备(如手机、平板电脑等移动设备)中部署，限制了其在实际生活的应用[14]。另外，Transformer模型最初为文本到文本的机器翻译任务设计，直接应用到手语翻译中存在模态障碍，即输入数据不再是文本而是视频序列，所以这种模态的跨越限制了模型语义特征的理解和上下文信息的提取能力。因此对多模态数据对齐研究的缺乏限制了进一步提升手语翻译质量。

针对上述存在的问题，本研究对于手语翻译的贡献如下：

(1) 提出基于Transformer模型的非自回归模型(Transformer Sign Language Translation Non-Autoregression, Trans-SLT-NA)，本模型利用自注意力机制强大的序列建模和上下文理解能力，并设计非自回归的文本生成算法进行手语翻译，改进了传统逐词生成的方式，一次性生成语句；

(2) 提出基于对比学习损失函数的视频-文本多模态数据对齐方法，通过约束视频和文本中间特征量的相似度为模型的文本提供指导，保证生成文本的正确性和流畅性；

(3) 在德语、中文和英文3个数据集上的验证结果表明本方法可以大幅度提升翻译速度，双语评估辅助指标(BiLingual Evaluation Understudy, BLEU-4)和自动摘要评估指标(Recall-Oriented Understudy for Gisting Evaluation, ROUGE)等文本生成质量指标接近自回归模型，可以更快速地减少模型训练过程中的复杂性。

2　相关工作

2.1　手语翻译任务

目前学术界对手语翻译任务的研究大多是逐词生成(one-by-one)的自回归翻译方式，该方式翻译质量高但速度较慢。Camgoz等人[15]最早采用2D-CNN框架进行空间嵌入，以及seq-to-seq的注意力模型进行序列映射，建立了端到端模型手语翻译模型和德语手语翻译数据集PHOENIX-2014T，被广泛使用并成为评价手语翻译模型的基准。Arvanitis等人[16]采用基于门控循环单元的seq-to-seq框架来进行手语翻译任务，使用3种不同的注意力机制来计算编码器和解码器的隐藏状态对齐权重参数，编码器-解码器系统具有较好的性能，但时间和空间开销大。而Transformer注意力模型计算开销较小且能提供更好的结果，因此可以通过该模型构建一个改进的、更可靠的手语翻译系统。

Xie等人[17]通过构建的内容感知和位置感知卷积层以及注入相对位置信息的编码器和解码器Transformer模型，在3个基准手语测试集上提高了1.6个BLEU。Chen等人[18]提出了一种用于手语翻

译的简单转移学习基线，使用手语到注释文本(sign-gloss)和注释文本到自然语言(gloss-text)任务对视觉和语言的联合模型进行预训练，然后通过视觉语言映射器的附加模块连接这两个网络进行微调，将二者结合起来实现手语翻译任务，进一步提升了翻译质量。Zhou等人[19]提出了back-translation手语反向翻译方法来解决平行手语文本数据有限的问题，通过设计回签翻译(sign back translation)传递，生成伪标签合成的平行数据并进行编码器-解码器手语翻译框架的端到端训练，将语言文本转换为源符号序列，同时将合成对作为额外的训练数据来处理，有效缓解了训练中并行数据短缺的问题。

目前基于自回归的手语翻译研究逐个生成单词的方式更加符合人类的直觉，因此具有较好的翻译质量，然而其缺点在于推理速度较慢，同时带来较大的计算开销。此外，基于Transformer的跨模态对齐约束可以有效改善手语翻译多模态任务特征交互的一致性[20]，对于提升手语翻译的质量较为关键。

2.2 非自回归机器翻译

传统的机器翻译模型框架为编码器-解码器模式，大多是基于循环神经网络(Recurrent Neural Network, RNN)或Transformer以自回归的形式进行翻译，后因翻译效率低从而使研究人员转向对非自回归模型的研究。Gu等人[21]首先提出了独立和同时生成目标令牌(target tokens)的非自回归机器翻译，与传统自回归模型相比，其采用类似的编码器-解码器框架，需要显式地预先定义或预测目标语句的长度，然后并行地生成目标语句的各个单词，可以极大提升翻译速度，但是存在着无法区分相邻的解码器隐藏状态，或者隐藏状态未完全传递源端信息而导致重复翻译和不完整翻译的问题[22]，在翻译质量上不及自回归模型，因此后续的研究都致力于提高非自回归翻译的质量。

Wang等人[22]通过引入在解码器输出中的相邻隐藏之间加入相似度约束，以及加入对偶学习的思想辅助正则项来改善非自回归翻译模型的解码器隐藏表示质量，并在多个基准数据集上验证了正则化策略有效性，提高了模型的准确性和推理效率。Xie等人[23]基于条件掩码语言模型框架引入了多视图子集正则化，通过预测目标句子中的随机掩码子集来训练条件翻译模型，该方法在WMT16 Ro-En和IWSLT14 De-En数据集上与更强的Transformer基线的差距缩小到0.01～0.44 BLEU分数。Zhou等人[24]提出了利用空时多线索网络来解决视觉的序列学习问题，采用多线索学习方法及联合优化策略和分段注意机制来充分利用多线索来源进行手语识别和翻译，在PHOENIX-2014,CSL和PHOENIX-2014T 3个大规模手语基准测试集上达到较高性能水平。

学者们基于非自回归机器翻译的编码器-解码器框架提出了不同的网络架构，可以消除解码器输入对先前目标符的依赖，并通过引入改进损失函数和解码算法以及利用预训练模型等方式来更好地捕捉目标依赖性。同时其具有更高的推理速度和处理吞吐量，也可以通过引入噪声或者利用自适应方法来处理错误，从而实现并行化计算，减少时间上的依赖性，提高长句子、复杂结构和多义词等手语翻译的准确性。

3 非自回归手语翻译模型

3.1 问题定义

输入手语视频$\boldsymbol{V}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}\in R^{T\times C\times H\times W}$，其中$\boldsymbol{x}_i$表示视频的第$i$帧图像，并基于Transformer获取时空特征序列，将序列经过位置编码后输入编码器中进行序列建模，然后完成手语视频到手语文本的映射，如图1所示。

其中，手语识别转换器(Sign Language Recognition Transformer, SLRT)目标是在学习有意义的

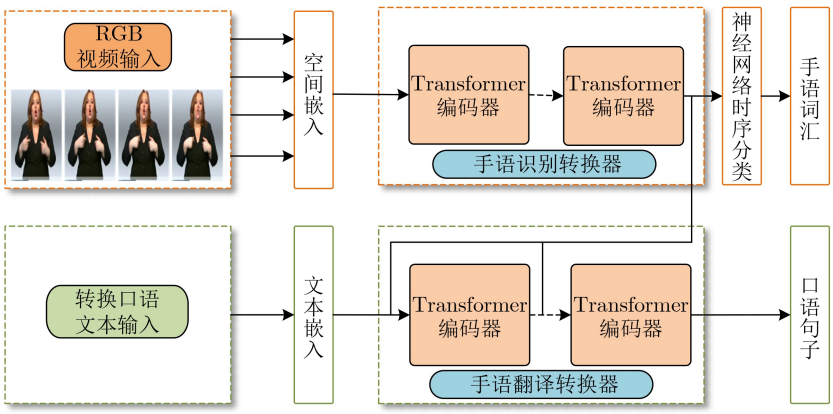


图 1 基于Transformer的连续手语识别和翻译框架

空间时间表征的同时，从连续的手语视频中识别光泽，以实现手语翻译。手语翻译转换器(Sign Language Translation Transformer, SLTT)则根据SLRT的标志性视频生成最后的口语句子，其在自注意力层输入上使用掩码，从SLRT和SLTT自注意力层提取的表示被组合并被提供给编码器解码器注意力模块，该模块学习源序列和目标序列之间的映射，但仍存在着显式跨模态对齐缺乏隐式自编码器对齐[20]。

待生成的目标序列$\boldsymbol{Y}=\{y_1,y_2,\cdots,y_N\}$表示与手语视频$\boldsymbol{V}$对应的文本，其中$y_i$表示自然语言语句中的第$i$个单词。在基于Transformer的连续手语识别和翻译的基础上，本文目标是设计深度学习模型F，使其通过视频-文本对$(\boldsymbol{V}^k,\boldsymbol{Y}^k)$数据的训练，学习手语和自然语言的对应关系并且能够具有较好的泛化能力，即

$$\hat{\boldsymbol{Y}}=F(\boldsymbol{V}|\theta) \tag{1}$$

其中，θ为模型的参数。为了获取更高质量的翻译文本，建立不同模态的数据约束关系，本文引入了对比学习损失函数用于视频序列数据与文本数据之间的对齐。基于上述内容，提出了非自回归手语翻译模型(Transformer Sign Language Translation Non-Autoregression, Trans-SLT-NA)，模型的整体结构如图2所示。

3.2 非自回归手语翻译模型Trans-SLT-NA

非自回归手语翻译模型主要由4个模块组成，分别是视频编码器、文本编码器、对比模块以及解码器。首先手语视频经过空间embedding转化为特征向量序列，然后视频编码器对序列进行编码，学习视频的上下文信息，并产生表征视频的对比嵌入向量。文本编码器在训练时用于对目标序列进行编码和上下文学习，并且获取表征文本语义信息的特殊嵌入向量。解码器对编码后的视频表征与文本特征使用交叉注意力机制进行交互，从而生成翻译后的文本。对齐模块则是将视频对比嵌入向量和文本特殊嵌入向量进行对齐处理，通过计算对比损失函数，将配对的向量距离拉近，不配对的向量距离远离，从而增加数据对齐约束。

(1) 视频编码器。视频编码器用于对手语视频进行编码并且学习上下文信息，将每一帧图像转化为包含上下文信息的特征向量。该模块由两部分组成，分别是由卷积神经网络组成的空间embedding部分和Transformer编码器组成的时序信息编码部分，如图3所示。

对于手语视频$\boldsymbol{V}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_T\}\in R^{T\times C\times H\times W}$，空间embedding层**SE**首先将其转化为向量序列

$$\boldsymbol{S}=\{\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T\}=\mathbf{SE}(\boldsymbol{V}) \tag{2}$$

其中，$\boldsymbol{s}_i=\mathbf{SE}(\boldsymbol{x}_i)\in R^{C'}$表示第$i$帧图像经过空间嵌入后的向量，$\boldsymbol{S}\in R^{T\times C'}$表示视频嵌入处理后的序列。在本研究中**SE**为预训练的EfficientNet-B0模型，用于提取图像的特征。

时序编码部分$\boldsymbol{f}_{\mathrm{v}}$由Transformer编码器组成，用于学习视频的上下文信息。由于本研究使用非自回归文本生成方式，因此需要在学习视频表示中训练模块用于预测目标序列的长度，所以在输入编码器的序列中加入表示目标序列长度的特殊向量[**len**]。具体计算过程为

$$\{\boldsymbol{p}_{\mathrm{len}},\ \boldsymbol{p},\boldsymbol{p}_{\mathrm{con}}\}=\boldsymbol{f}_{\mathrm{v}}(\{[\mathbf{len}],\boldsymbol{s}_1,\boldsymbol{s}_2,\cdots,\boldsymbol{s}_T,[\mathbf{con}]\}) \tag{3}$$

其中，$\boldsymbol{P}=\{\boldsymbol{p}_1,\boldsymbol{p}_2,\cdots,\boldsymbol{p}_T\}\in R^{T\times C'}$为编码后的视

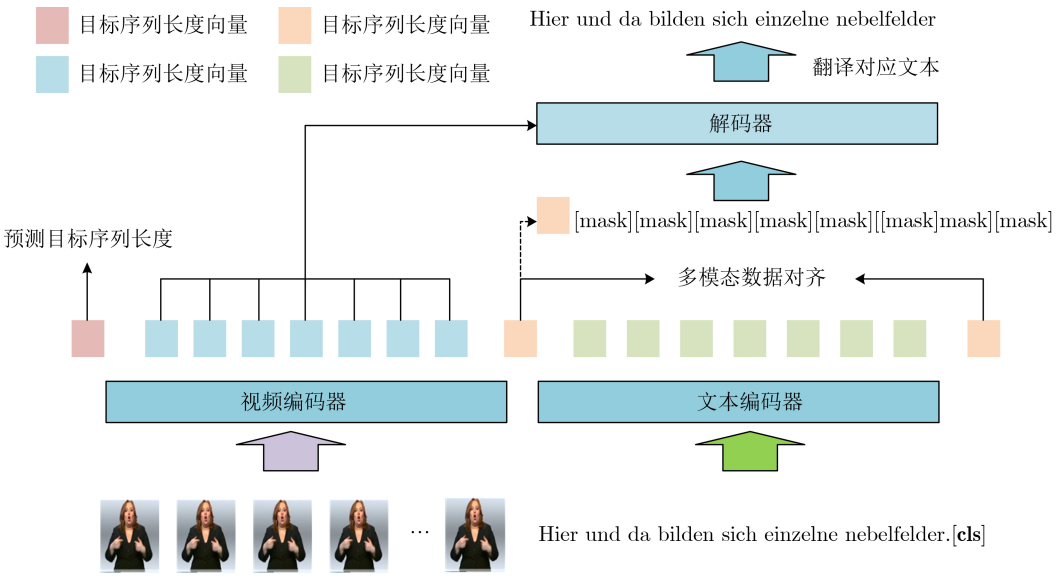


图 2 Trans-SLT-NA模型总体结构图

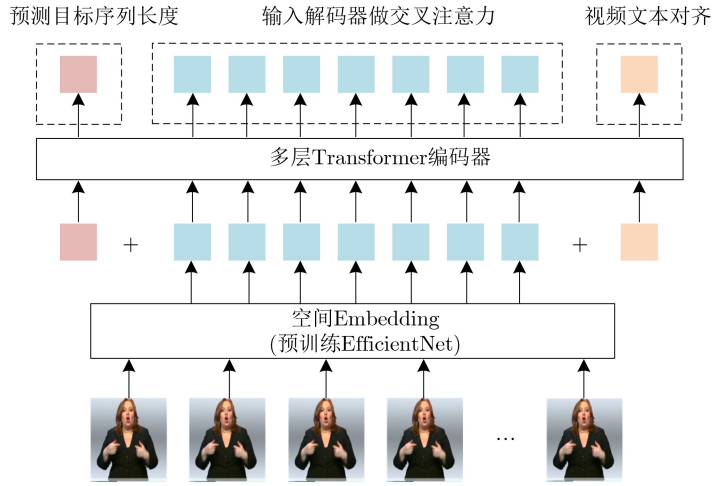


图 3 视频编码器的组成结构

频特征向量，$\boldsymbol{p}_{\text{len}}$用于预测目标序列的长度，$\boldsymbol{p}_{\text{con}}$用于计算对比损失进行视频文本对齐。时序编码部分的模型结构为Transformer的编码器，其核心为注意力机制，其具体计算为

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{C'}}\right)\boldsymbol{V} \tag{4}$$

其中，$\boldsymbol{Q}$, $\boldsymbol{K}$和$\boldsymbol{V}$分别表示Query, Key和Value，均由输入序列产生，用于计算序列中每一个元素与其他元素的相关性并加权求和。

(2) 文本编码器。与视频编码器类似，文本编码器由词语embedding层和Transformer编码器部分组成，词语embedding层$\mathbf{WE}$用于将单词转化为计算机可以识别处理的词向量，然后Transformer编码器$\boldsymbol{f}_{\text{w}}$对词向量序列学习时序表达和提取上下文信息。文本编码器模块主要感知文本的语义信息，产生文本的特征向量用于计算对比损失，具体计算为

$$\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N\}=\mathbf{WE}\left(\boldsymbol{Y}\right) \tag{5}$$

$$\{\boldsymbol{Z},\ \boldsymbol{z}_{\text{cls}}\}=\boldsymbol{f}_{\text{w}}\left(\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_N,[\mathbf{cls}]\}\right) \tag{6}$$

其中，$\boldsymbol{Z}$表示经过编码器编码后的文本的向量，[**cls**]表示添加在序列末尾的特殊嵌入向量，$\boldsymbol{z}_{\text{cls}}$表示添加的[**cls**]编码后通过自注意力机制和Transformer编码器模型处理得到的特征向量，可以捕捉文本的全局信息和高级语义信息，用于视频-文本的对齐。

(3) 对比模块。多模态学习中，对比学习用于拉近视频模态和对应文本模态的特征距离，使两种模态的表示在特征空间中更加接近。对比模块用于将多模态的视频和文本数据进行对齐处理，在本模块中，设计了对比损失函数来约束视频和文本的对齐关系。

在本文的对比模块中，视频-文本对齐作用是以手语视频特征为中心，将与之匹配的文本特征拉近，反之将与之不匹配的文本特征拉远。与此类似，在文本-视频对齐中，是以文本特征为中心进行计算。视频-文本对齐、文本-视频对齐二者的核心思想一致，区别在于视频-文本对齐中以视频为中心计算，在文本-视频对齐中以文本为中心计算，如此可以强化跨模态数据的对齐学习。

该模块的核心思想是对于一个配对的手语视频和文本$(\boldsymbol{V},\boldsymbol{Y})$，经过视频编码器和文本编码器的处理后分别得到表征视频的向量$\boldsymbol{p}_{\text{con}}$和表征文本的向量$\boldsymbol{z}_{\text{cls}}$，由于这两者是配对的数据，则意味着其在语义上是相近的，因此两个向量的距离应当尽可能靠近，相反对于不配对的视频和文本其向量应该远离。因此对比损失设计为

$$L_{\text{con}}=-\frac{1}{B}\left(\underbrace{\sum_{i}^{B}\ln\frac{\exp\left(\frac{\boldsymbol{p}_i\cdot\boldsymbol{z}_i}{\sigma}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\boldsymbol{p}_i\cdot\boldsymbol{z}_j}{\sigma}\right)}}_{\text{视频-文本对齐}}+\underbrace{\sum_{i}^{B}\ln\frac{\exp\left(\frac{\boldsymbol{z}_i\cdot\boldsymbol{p}_i}{\sigma}\right)}{\sum_{j=1}^{B}\exp\left(\frac{\boldsymbol{z}_i\cdot\boldsymbol{p}_j}{\sigma}\right)}}_{\text{文本-视频对齐}}\right) \tag{7}$$

$$\boldsymbol{p}=\text{Norm}\left(\boldsymbol{p}_{\text{con}}\right),\ \boldsymbol{z}=\text{Norm}\left(\boldsymbol{z}_{\text{cls}}\right) \tag{8}$$

其中，B表示批量数据中包含的样本量，$\boldsymbol{p}_i$和$\boldsymbol{z}_i$表示标准化后的第i个配对视频和文本的表征向量，根据损失函数可以看出，每一项的分子表示配对的向量靠近，分母表示不配对的向量远离，从而实现多模态数据的对齐。

(4) 解码器。解码器用于根据视频编码器的输出从而生成对应的文本，该模块由Transformer的解码器构成，在生成策略上与传统的Transformer不同，传统方法是自回归的方法，即模型1次运算只生成1个词语，然后按照顺序依次生成，直到终止符为止。本文的算法无需模型多次运行逐词生成，只需1次运行即可并行地生成完整语句。首先根据视频编码器的输出预测目标序列长度M，然后使用特殊字符[mask]初始化序列，即得到初始化后的待生成的文本序列$\boldsymbol{O}=\{[\text{mask}]_1,[\text{mask}]_2,\cdots,[\text{mask}]_M\}$，在输入解码器$\boldsymbol{f}_{\text{D}}$时，在序列前加入视频表征向量$\boldsymbol{p}_{\text{con}}$作为引导信息，文本生成过程为

$$\hat{\boldsymbol{Y}}=\{\hat{y}_1,\hat{y}_2,\cdots,\hat{y}_M\}=\boldsymbol{f}_{\text{D}}\left(\boldsymbol{p}_{\text{con}},\boldsymbol{O}|\boldsymbol{P}\right) \tag{9}$$

其中，$\hat{\boldsymbol{Y}}$表示解码器根据视频生成对应的文本，M表示根据$\boldsymbol{p}_{\text{len}}$确定的序列长度，另外$\boldsymbol{P}$和$\boldsymbol{p}_{\text{len}}$均由式(3)计算得到。该模块的核心仍然是注意力机制，区别在于解码过程需要用到视频信息，因此采用的交叉注意力机制而非自注意力，即式(4)中的$\boldsymbol{Q}$由文本序列$\boldsymbol{O}$产生，而$\boldsymbol{K}$和$\boldsymbol{V}$均由视频特征$\boldsymbol{P}$得到。

3.3 模型训练

本文使用手语数据集对模型进行训练，其中视频编码模块需要被训练理解手语视频的表示，文本编码模块需要被训练获取文本的语义信息，解码器需要准确地将对应的词语进行预测，此外还需要对目标序列长度进行预测。

首先，对于解码器预测的单词，本文使用交叉熵损失函数来量化标签与目标的损失

$$L_{\text{pred}}=-\sum_{i=1}^{M}\ln P_\theta\left(\hat{y}_i|\boldsymbol{P}\right) \tag{10}$$

对于目标序列的长度预测，使用交叉熵损失函数来计算误差

$$L_{\text{len}}=-\ln P_\theta\left(\boldsymbol{p}_{\text{len}}\right) \tag{11}$$

最终，完整的损失计算由3部分组成

$$L=\lambda_{\text{c}}\cdot L_{\text{con}}+\lambda_{\text{p}}\cdot L_{\text{pred}}+\lambda_{\text{l}}\cdot L_{\text{len}} \tag{12}$$

其中，λ_{c}，λ_{p}和λ_{l}为超参数，用于平衡各项损失函数的值，λ_{l}设置为0.01，并且约束$\lambda_{\text{c}}+\lambda_{\text{p}}=1$，这是因为在实验中发现长度预测的loss较小并且容易过拟合，因此将参数设置较小增加其反向传播的梯度约束。训练算法采用梯度下降算法通过最小化损失函数的值来迭代更新模型参数从而使模型具有理解手语并翻译为自然语言的能力。

4 实验分析

4.1 实验数据

选取手语翻译研究中被广泛应用的德语数据集PHOENIX-2014T[15]、中文手语数据集CSL-Daily[19]和英文手语数据集How2Sign[25]来评估所提出的模型和算法。PHOENIX-2014T数据集为不同手语演示者的天气预报广播视频片段，并且对连续片段进行标注，该数据集共包括8 257 条视频数据，其中7 096 条视频为训练集，验证集和测试集各有519和642 条。CSL-Daily数据集是在室内录制的中文手语数据集，数据集包含20 654 条视频和对应文本的数据，由10 个不同的手语演示者展示，包含了学校、生活、医疗等多种场景，该数据集被划分为18 401, 1 077和1 176被用于训练、验证和测试。How2Sign数据集由11 名手语演示者在绿布背景下进行手语表演，其中训练集有31 128 条视频，验证集有1 741 条视频，测试集有2 322 条视频，涵盖16 000 个词汇量。表1给出了3种数据集所对应的信息。

4.2 参数设置及评估指标

对于本文所提出的Trans-SLT-NA模型将视频编码器、文本编码器和解码的Transformer编码器或解码器层数设为3 层，每一层的注意力机制的头数设置为16，隐藏层的向量维度设为1 024，该模型的尺寸与后续对比文献中设置相同便于比较结果。

所有代码均使用Python 3.9和Pytorch 1.10进行编写，初始学习率设置为5e–4，并且使用warm-up和逐步降低学习率策略，使用Adam优化器且参数均使用默认设置，使用批量数据训练，batch size设为16。实验环境使用2块Nvidia RTX 3090 GPU进行加速训练。

对于翻译质量的评价，选择使用机器翻译中广泛使用的指标BLEU-4[5,24]，另外使用的指标为ROUGE指标，该指标主要衡量生成文本与参考文本的重叠程度[15]。

4.3 对比SOTA结果

首先在德语数据集PHOENIX-2014T上与目前的各种方法的对比结果如表2所示。根据实验结

表 1　训练模型使用的数据集信息

数据集	语言	训练集	验证集	测试集	总数
PHOENIX-2014T	德语	7 096	519	642	8 257
CSL-Daily	中文	18 401	1 077	1 176	20 654
How2Sign	英文	31 128	1 741	2 322	35 191

果，首先是在推理速度上，以同样是Transformer架构的模型SLTR-T为基准，本文提出的Trans-SLT-NA模型的推理速度提升了11.6 倍，相比于其他自回归方式的手语翻译方法其推理速度也显著提升(表中无速度对比的原因是相关研究未开源代码和模型，表3同理)。从翻译质量上看，本文提出的模型在BLEU-4和ROUGE指标上均要显著优于RNN-based模型，由此可以看出自注意力机制在对序列数据的理解上要优于循环神经网络。与此同时，Trans-SLT-NA在指标上虽然略低于其他的自回归模型，但是非自回归由于一次生成的策略，在生成质量方面不如自回归模型，原因是自回归模型在每一步都可以使用之前时刻生成的语句，可利用的信息较强，这一点在机器翻译任务上已经获得学界广泛认可。

在中文手语数据集CSL-Daily上的评估结果如表3所示，由于使用了一次性并行生成的方式，本文提出的方法相比较于自回归模型推理速度提升了13.4 倍。同时Trans-SLT-NA模型在BLEU-4和ROUGE指标上表现优于SLTR-T和ConSLT两个基于Transformer的自回归模型，值得一提的是Contrastive-T模型也引入了对比学习的技术，区别在于其并没有进行视频和文本数据的对齐。

在英文手语数据集How2Sign上的评估结果如表4所示，本文所提出的非自回归的方法与当前的基线方法对比，在测试集上的BLEU-4指标要高出0.55，在翻译质量上两者几乎相同，但是在推理速度上本文所提出的非自回归方法要快17.6 倍。进一步分析，推理速度提升的原因是因为该数据集的文本长度较长(相比较德语和中文的数据集)，文本长度越长，自回归方式就越慢，而非自回归策略在速度上几乎不受影响，因此在长文本翻译中，非自回归的优势将更加显著。

上述3个数据集的评估结果表明，本文使用的非自回归的方式进行手语翻译，不需要端到端的训练，可以节省计算资源，在翻译速度上相较于自回归模型显著。同时在翻译质量上也与自回归模型接近，表明模型可以更好地捕捉手语和文本之间的内在联系，证明了本方法的有效性。

4.4 消融实验

消融实验包括多模态数据对齐的效果评估、空间embedding对翻译质量的影响研究和损失函数的

表 2 模型在PHOENIX-2014T数据集上的结果

方法	生成方式	验证集		测试集		推理速度
		BLEU-4	ROUGE	BLEU-4	ROUGE	
RNN-based[15]	AR	9.94	31.8	9.58	31.8	2.3X
SLTR-T[5]	AR	20.69	–	20.17	–	1.0X
Multi-C[26]	AR	19.51	44.59	18.51	43.57	–
STMC-T[24]	AR	24.09	48.24	23.65	46.65	–
PiSLTRc[17]	AR	21.48	47.89	21.29	48.13	0.92X
Trans-SLT-NA	NAR	18.81	47.32	19.03	48.22	11.6X

注：AR表示自回归生成方式，NAR表示非自回归生成。

表 3 CSL-Daily数据集上的对比结果

方法	生成方式	验证集		测试集		推理速度
		BLEU-4	ROUGE	BLEU-4	ROUGE	
SLTR-T[5]	AR	11.88	37.06	11.79	36.74	1X
Sign Back-Tran[19]	AR	20.80	49.49	21.34	49.31	0.89X
ConSLT[27]	AR	14.80	41.46	14.53	40.98	–
Trans-SLT-NA	NAR	16.22	43.74	16.72	44.67	13.4X

表 4 How2Sign数据集上的对比结果

方法	生成方式	验证集		测试集		推理速度
		BLEU-4	ROUGE	BLEU-4	ROUGE	
Baseline	AR	8.89	–	8.03	–	1X
Trans-SLT-NA	NAR	8.14	32.84	8.58	33.17	17.6X

超参数影响研究。本文引入了对比学习损失函数进行多模态数据的对齐，从而在浅空间中将视频数据特征和对应的文本语义信息拉近，实验验证结果如表5所示。消融实验结果表明，在德语数据集的BLEU-4得分在验证集和测试集中分别提升了2.79和3.06分，在中文手语数据集中分别提升了1.79和1.51分，在英文手语数据集中分别提升了0.33和0.35分。表明引入多模态对齐机制可以使手语翻译的结果更加流畅，与实际的文本更接近，对比学习损失函数进行多模态数据视频-文本的对齐是有效的，在3种手语数据集上使用数据对齐策略后均能使翻译质量提升。

为了对比多模态数据对齐的效果，使用t-SNE方法将视频表征向量和文本表征向量进行降维可视化，从而直观地对比向量距离，结果如图4所示。结果表明，不使用数据对齐可使视频表征向量之间的距离更远，视频数据之间以及文本数据之间的相似度更低，模型并没有学习到潜空间的语义对应关系。使用数据对齐后明显看出对应的视频和文本的距离更接近，说明模型已经学习到视频与文本之间的语义对应关系。

空间embedding旨在将彩色图像经过卷积操作提取其空间特征从而转化为特征向量，因此能否充分理解图像并提取信息对于翻译的准确性起到关键作用，本研究采用的空间embedding为预训练EfficientNet-B0。为了研究该模块对手语翻译质量的影响，设计消融实验研究不同卷积网络以及预训练对最终手语翻译准确性和流畅度的效果进行验证，其中预训练表明卷积网络在ImageNet数据集上进行训练，实验结果如表6所示。

从表6可以看出，使用经过预训练的网络翻译质量更高，说明其对于图像的语义信息提取更为充分，BLEU-4指标平均提升2.38，反映出翻译准确性和流畅度较高，横向比较来看，EfficientNet-B0效果要优于ResNet-50和VGG-19。实验结果表明图像语义信息对于手语翻译的质量有影响，充分地提取图像特征有利于手语视频的理解。

在3.3节中对3部分损失函数设置了系数，其中λ_l固定为0.01，另外两个参数的变化对模型性能的影响，本文设计消融实验进行验证结果如表7所示。结果表明不使用对比损失函数时(即不进行数据对齐)模型效果最差，该结论与消融实验部分一致，另外当λ_c与λ_p相等时模型的效果达到最优，因此本实验选取$\lambda_c=\lambda_p=0.5$的参数组合。该消融实验说明数据对齐有利于提升模型手语翻译的质量，如果数据对齐的权重较高，则不利于模型解码器对于单词的生成，因此需要合理调整权重。

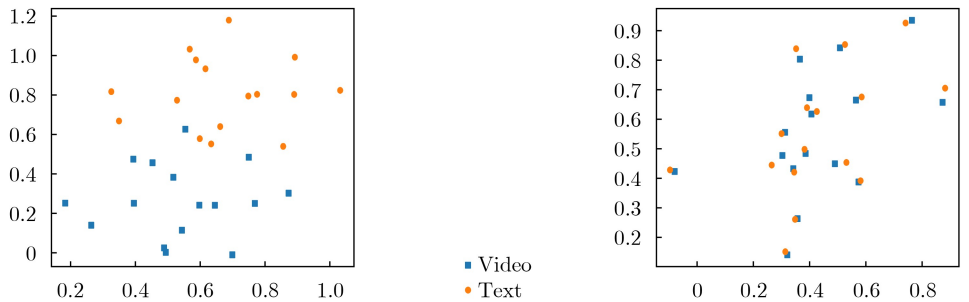
5　结束语

本文针对当前自回归式的手语翻译速度缓慢的

表 5　多模态数据对齐的有效性验证

模型	数据集	数据对齐	验证集		测试集	
			BLEU-4	ROUGE	BLEU-4	ROUGE
Trans-SLT-NA	PHOENIX-2014T	w	18.81	47.32	19.03	48.22
		w/o	16.02	43.21	15.97	42.85
	CSL-Daily	w	16.22	43.74	16.72	44.67
		w/o	14.43	42.27	15.21	42.84
	How2Sign	w	8.14	32.84	8.58	33.17
		w/o	7.81	30.16	8.23	30.59

注：w表示使用数据对齐，w/o表示不使用数据对齐。



(a) 未使用多模态数据对齐　　(b) 使用多模态数据对齐

图 4 使用t-SNE对视频表征向量和文本向量的可视化

表 6 空间Embedding对于模型性能的影响结果

空间Embedding	预训练	验证集		测试集	
		BLEU-4	ROUGE	BLEU-4	ROUGE
VGG-19	w/o	14.42	38.76	14.36	39.17
ResNet-50	w/o	15.57	40.26	15.33	41.17
EfficientNet-B0	w/o	16.32	40.11	16.04	41.27
VGG-19	w	16.84	43.31	16.17	42.09
ResNet-50	w	17.79	45.63	16.93	44.53
EfficientNet-B0	w	18.81	47.32	19.03	48.22

表 7 损失函数超参数对于模型性能的结果

λ_p	λ_c	验证集		测试集	
		BLEU-4	ROUGE	BLEU-4	ROUGE
1	0	16.02	43.21	15.97	42.85
0.8	0.2	17.37	44.89	16.87	42.46
0.5	0.5	18.81	47.32	19.03	48.22
0.2	0.8	18.04	46.17	18.26	47.10

问题，研究了非自回归手语翻译，以及多模态数据交互性较差导致翻译质量不佳现象。基于自注意力模型Transformer以及对比损失函数设计了带有多模态数据对齐的非自回归手语翻译模型Trans-SLT-NA，该模型根据输入的手语视频预测目标文本的长度，然后通过解码器并行地预测所有的单词，从而极大地加速了推理过程。此外为模型增加了视频-文本的多模态对齐模块，将对应的视频和文本特征的空间距离拉近，从而增强模型的表达能力。本研究可以为基于深度学习手语翻译模型的应用与部署提供基础，未来可以利用知识蒸馏等压缩量化技术进一步优化模型的复杂度，实现技术的落地与应用。

参 考 文 献

[1] 闫思伊, 薛万利, 袁甜甜. 手语识别与翻译综述[J]. 计算机科学与探索, 2022, 16(11): 2415–2429. doi: 10.3778/j.issn.1673-9418.2205003.
YAN Siyi, XUE Wanli, and YUAN Tiantian. Survey of sign language recognition and translation[J]. *Journal of Frontiers of Computer Science and Technology*, 2022, 16(11): 2415–2429. doi: 10.3778/j.issn.1673-9418.2205003.

[2] 陶唐飞, 刘天宇. 基于手语表达内容与表达特征的手语识别技术综述[J]. 电子与信息学报, 2023, 45(10): 3439–3457. doi: 10.11999/JEIT221051.
TAO Tangfei and LIU Tianyu. A survey of sign language recognition technology based on sign language expression content and expression characteristics[J]. *Journal of Electronics & Information Technology*, 2023, 45(10): 3439–3457. doi: 10.11999/JEIT221051.

[3] DUARTE A, PALASKAR S, VENTURA L, *et al*. How2Sign: A large-scale multimodal dataset for continuous American sign language[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 2734–2743. doi: 10.1109/CVPR46437.2021.00276.

[4] 周乐员, 张剑华, 袁甜甜, 等. 多层注意力机制融合的序列到序列中国连续手语识别和翻译[J]. 计算机科学, 2022, 49(9): 155–161. doi: 10.11896/jsjkx.210800026.
ZHOU Leyuan, ZHANG Jianhua, YUAN Tiantian, *et al*. Sequence-to-sequence Chinese continuous sign language recognition and translation with multilayer attention mechanism fusion[J]. *Computer Science*, 2022, 49(9): 155–161. doi: 10.11896/jsjkx.210800026.

[5] CAMGÖZ N C, KOLLER O, HADFIELD S, *et al*. Sign language transformers: Joint end-to-end sign language recognition and translation[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, USA, 2020: 10020–10030. doi: 10.1109/CVPR42600.2020.01004.

[6] HUANG Jie, ZHOU Wengang, ZHANG Qilin, *et al*. Video-based sign language recognition without temporal segmentation[C]. 32nd AAAI Conference on Artificial Intelligence, New Orleans, USA, 2018: 2257–2264. doi: 10.1609/aaai.v32i1.11903.

[7] ZHOU Hao, ZHOU Wengang, and LI Houqiang. Dynamic pseudo label decoding for continuous sign language recognition[C]. 2019 IEEE International Conference on Multimedia and Expo (ICME), Shanghai, China, 2019: 1282–1287. doi: 10.1109/ICME.2019.00223.

[8] SONG Peipei, GUO Dan, XIN Haoran, *et al*. Parallel temporal encoder for sign language translation[C]. 2019 IEEE International Conference on Image Processing (ICIP), Taipei, China, 2019: 1915–1919. doi: 10.1109/ICIP.2019.8803123.

[9] VASWANI A, SHAZEER N, PARMAR N, *et al*. Attention is all you need[C]. The 31st International Conference on Neural Information Processing Systems, Long Beach, USA, 2017: 6000–6010.

[10] 路飞, 韩祥祖, 程显鹏, 等. 基于轻量3D CNNs和Transformer的手语识别[J]. 华中科技大学学报:自然科学版, 2023, 51(5): 13–18. doi: 10.13245/j.hust.230503.
LU Fei, HAN Xiangzu, CHENG Xianpeng, *et al*. Sign language recognition based on lightweight 3D CNNs and transformer[J]. *Journal of Huazhong University of Science and Technology:Natural Science Edition*, 2023, 51(5): 13–18. doi: 10.13245/j.hust.230503.

[11] WANG Hongyu, MA Shuming, DONG Li, *et al*. DeepNet: Scaling transformers to 1, 000 layers[EB/OL].

https://arxiv.org/abs/2203.00555, 2022.

[12] KISHORE P V V, KUMAR D A, SASTRY A S C S, *et al*. Motionlets matching with adaptive kernels for 3-D Indian sign language recognition[J]. *IEEE Sensors Journal*, 2018, 18(8): 3327–3337. doi: 10.1109/JSEN.2018.2810449.

[13] XIAO Yisheng, WU Lijun, GUO Junliang, *et al*. A survey on non-autoregressive generation for neural machine translation and beyond[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(10): 11407–11427. doi: 10.1109/TPAMI.2023.3277122.

[14] LI Feng, CHEN Jingxian, and ZHANG Xuejun. A survey of non-autoregressive neural machine translation[J]. *Electronics*, 2023, 12(13): 2980. doi: 3390/electronics 12132980.

[15] CAMGOZ N C, HADFIELD S, KOLLER O, *et al*. Neural sign language translation[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 7784–7793. doi: 10.1109/CVPR.2018.00812.

[16] ARVANITIS N, CONSTANTINOPOULOS C, and KOSMOPOULOS D. Translation of sign language glosses to text using sequence-to-sequence attention models[C]. 2019 15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Sorrento, Italy, 2019: 296–302. doi: 10.1109/SITIS.2019.00056.

[17] XIE Pan, ZHAO Mengyi, and HU Xiaohui. PiSLTRc: Position-informed sign language transformer with content-aware convolution[J]. *IEEE Transactions on Multimedia*, 2022, 24: 3908–3919. doi: 10.1109/TMM.2021.3109665.

[18] CHEN Yutong, WEI Fangyun, SUN Xiao, *et al*. A simple multi-modality transfer learning baseline for sign language translation[C]. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, USA, 2022: 5110–5120. doi: 10.1109/CVPR52688.2022.00506.

[19] ZHOU Hao, ZHOU Wengang, QI Weizhen, *et al*. Improving sign language translation with monolingual data by sign back-translation[C]. 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, USA, 2021: 1316–1325. doi: 10.1109/CVPR46437.2021.00137.

[20] ZHENG Jiangbin, WANG Yile, TAN Cheng, *et al*. CVT-SLR: Contrastive visual-textual transformation for sign language recognition with variational alignmen[C]. 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 23141–23150. doi: 10.1109/CVPR52729.2023.02216.

[21] GU Jiatao, BRADBURY J, XIONG Caiming, *et al*. Non-autoregressive neural machine translation[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018. doi: 10.48550/arXiv.1711.02281.

[22] WANG Yiren, TIAN Fei, HE Di, *et al*. Non-autoregressive machine translation with auxiliary regularization[C]. The 33rd AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 5377–5384. doi: 10.1609/aaai.v33i01.33015377.

[23] XIE Pan, LI Zexian, ZHAO Zheng, *et al*. MvSR-NAT: Multi-view subset regularization for non-autoregressive machine translation[J]. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2022: 1–10. doi: 10.1109/TASLP.2022.3221043.

[24] ZHOU HAO, ZHOU Wengang, ZHOU Yun, *et al*. Spatial-temporal multi-cue network for sign language recognition and translation[J]. *IEEE Transactions on Multimedia*, 2022, 24: 768–779. doi: 10.1109/TMM.2021.3059098.

[25] TARRÉS L, GÁLLEGO G I, DUARTE A, *et al*. Sign language translation from instructional videos[C]. The IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Vancouver, Canada, 2023: 5625–5635. doi: 10.1109/CVPRW59228.2023.00596.

[26] CAMGOZ N C, KOLLER O, HADFIELD S, *et al*. Multi-channel transformers for multi-articulatory sign language translation[C]. ECCV 2020 Workshops on Computer Vision, Glasgow, UK, 2020: 301–319. doi: 10.1007/978-3-030-66823-5_18.

[27] FU Biao, YE Peigen, ZHANG Liang, *et al*. A token-level contrastive framework for sign language translation[C]. 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023: 1–5. doi: 10.1109/ICASSP49357.2023.10095466.

邵舒羽：男，副教授，研究方向为信号处理、复杂系统可靠性分析.
杜　垚：男，博士生，研究方向为模式识别.
范晓丽：女，高级工程师，研究方向为生物医学信号处理、模式识别.

责任编辑：余　蓉