手语制作：一个回顾

Razieh Rastgoo[1,2] Kourosh基亚尼[1]，塞尔吉奥·埃斯卡莱拉[3]穆罕默德·萨博克鲁[2] [1]森南大学[2]基础科学研究所(IPM)
[3]巴塞罗那大学和计算机视觉中心
rrastgoo@semnan.ac.埃尔，库罗什。kiani@semnan.ac.ir, sergio@maia.ub.es, sabokro@ipm.ir

摘要

手语是聋人和听力障碍社区使用的主要但非主要的交流语言形式。为了使听力障碍群体和听力群体之间轻松相互沟通，建立一个能够将口语翻译成手语的强大系统，反之亦然是最基本的。为此，手语识别和生产是建立这种双向系统的两个必要部分。手语识别和生产需要应对一些关键的挑战。在这项调查中，我们回顾了使用深度学习在手语生产(SLP)和相关领域的最新进展。本调查旨在简要总结SLP的最新进展，讨论它们的优点、局限性和未来的研究方向。

1.介绍

手语是社会中大量人群中使用的主要但非主要的交流语言形式。根据世界卫生组织(WHO)2020年的报告，世界上有超过4.66亿聋人，[88]。不同民族的手语有不同的形式，如美国[87]，阿根廷[26]，波兰[71]，德国[36]，希腊[27]，西班牙[3]，China[2]，韩国[61]，Iran[28]，等等。为了使听力障碍群体和听力障碍群体之间轻松地相互沟通，建立一个能够将口语翻译成手语的强大系统，反之亦然是最基本的。为此，手语识别和生产是建立这种双向系统的两个必要部分。虽然第一部分，手语识别，近年来迅速发展，[64,65,66,67,68,69,50,62,59,8,48]，最新的一个，手语生产(SLP)，仍然是一个非常具有挑战性的问题，涉及到视觉和语言信息之间的解释[79]。所提出的符号识别系统通常会将符号映射到spo-中肯语言形式的文本转录[69]。然而，SLP系统会执行相反的过程。

手语识别和生产正在应对[69,79]面临的一些关键挑战。其中之一是符号的视觉变异性，它受到手的形状、手掌方向、运动、位置、面部表情和其他非手信号的影响。这些符号外观的差异产生了较大的类内变异性和较低的类间变异性。这使得很难提供一个能够识别不同符号类型的健壮的通用系统。另一个挑战是开发一个逼真的SLP系统，在现实世界中从文本或语音中生成相应的符号数字、单词或句子。与手语的语法规则和语言结构相对应的挑战是这一领域的另一个关键挑战。口语和手语之间的翻译是一个复杂的问题。这不是一个简单的从文本/语音到符号的映射问题。这一挑战来自于口语和手语中单词的标记化和顺序之间的差异。

另一个挑战与应用程序领域有关。大多数的应用程序在手语专注于手语识别如机器人[20]，人机交互[5]，教育[19]，电脑游戏[70]，识别自闭症儿童[11]，自动手语解释[90]，决定支持医学诊断运动技能障碍[10]，家庭康复[17][57]，和虚拟现实[82]。这是由于听力正常的人认为聋人更喜欢阅读口语；因此，没有必要把阅读的口语翻译成手语。这不是真的，因为不能保证一个聋人熟悉一种说话语言的阅读和写作形式。在某些语言中，这两种形式彼此之间完全不同。虽然在手语识别[30,69]中有一些详细而详尽的评论，但SLP却有如此详细的评论。在这里，我们提出了一个调查，包括最近在SLP的工作，目的是讨论

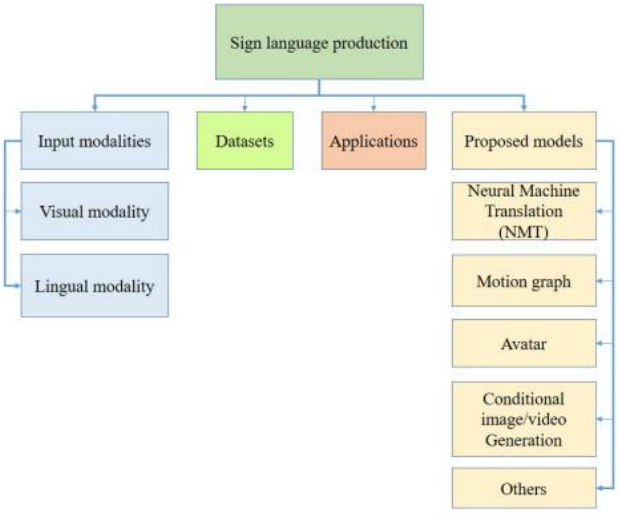

图1。在SLP中回顾的工作提出的分类。

这一领域的进步和弱点。我们专注于基于深度学习的模型来分析最先进的SLP技术。本文的其余部分组织如下。第2节介绍了一个分类法，它总结了与SLP相关的主要概念。最后，第3节讨论了SLP的发展、优势和局限性，并对未来研究的可能路线进行了评论。

2. SLP分类法

在本节中，我们将介绍一个分类法，它总结了与SLP中的深度学习相关的主要概念。我们对SLP中最近的工作进行了分类，并在每个类别中提供了单独的讨论。在本节的其余部分中，我们将解释不同的输入模式、数据集、应用程序和所提出的模型。图1显示了本节中描述的建议的分类法。

2.1. 输入方式

一般来说，视觉和语言是SLP中的两种输入模式。视觉模态包括捕获的图像/视频数据，而口语的语言模态包含来自自然语言输入的文本。计算机视觉和自然语言处理技术是处理这些输入模式的必要条件。视觉形态：RGB和骨架是SLP模型中使用的两种常见的输入数据类型。虽然RGB图像/视频包含高分辨率的内容，但骨架输入减少了输入模型所需的输入维度，并帮助制作一个低复杂和快速的模型。RGB图像输入中只包含一个字母或数字。输入图像对应的空间特征可以通过基于计算机视觉的技术来提取，特别是基于深度学习的模型。近年来，卷积神经网络(CNN)在输入图像[53]的空间特征提取方面取得了出色的性能。此外，生成模型，如生成对抗网络(GAN)，可以使用CNN作为一个编码器或解码器块来生成一个符号图像/视频。由于RGB视频输入的时间维度，这种输入模态的处理比RGB图像输入更为复杂。在SLP中提出的大多数模型都使用RGB视频作为输入[13,72,73,79]。一个RGB符号视频可以对应一个符号字或一些连接的符号字，以符号句的形式出现。GAN和LSTM是SLP中最常用于静态和动态视觉模式的基于深度学习的模型。虽然使用这些模型已经取得了成功的结果，但还需要更多的努力来生成更逼真的符号图像/视频，以改善与聋人社区的沟通界面。

语言形态：文本输入是语言形态中最常见的一种形式。为了处理输入文本，使用了不同的模型来[76,80]。与图像/视频处理相比，文本处理比较复杂，但文本翻译任务却比较复杂。在基于深度学习的模型中，神经机器翻译(NMT)模型是输入文本处理中最常用的模型。其他的Seq2Seq模型[80]，如基于递归神经网络(RNN)的模型，在许多任务中证明了它们的有效性。虽然使用这些模型取得了成功的结果，但需要更多的努力来克服翻译任务中现有的挑战。翻译中的挑战之一是由于不同语言中不同的词汇风格、翻译和意义而导致的领域适应。因此，开发机器翻译系统的一个关键要求是针对一个特定的领域。迁移学习，在一般领域中训练翻译系统，然后在几个时期对领域内数据进行ine调优，是应对这一挑战的一种常见方法。另一个挑战是关于训练数据的数量。由于基于深度学习的模型的一个主要特性是数据量和模型性能之间的相互关系，因此需要大量的数据来提供良好的泛化。另一个挑战是机器翻译系统在不常见和看不见单词上的性能不佳。为了处理这些单词，字节对编码，如翻译或复合分裂，可以用于罕见的单词翻译。另一个挑战是，机器翻译系统不能正确地翻译长句子。然而，注意力模型[86]部分地处理了短句的这一挑战。此外，关于单词对齐的挑战在反向翻译中更为关键，即从目标语言翻译回源语言。

.2.2个数据集

虽然有一些大规模的注释数据集可用于手语识别，但只有少数公开的可用的SLP数据集。两个公共数据集，RWTH-凤凰城-2014T[14]和How2Sign[22]是在手语翻译中使用最多的数据集。前者包括可用于文本到手语翻译的德语手语句子。该数据集是连续符号识别数据集的扩展版本，凤凰城-2014[29]。RWTH-PHOENIX天气2014T包含了由9个签名者执行的总共8257个序列。该数据集中有1066个符号注释和2887个口语词汇表。此外，在该数据集中还包含了与口语句子对应的光泽度注释。后面的数据集，How2Sing，是最近提出的一个用于语音到符号的语言翻译的多模态数据集。该数据集总共包含38611个序列和4k个由10个签名者执行的词汇表。与前一个数据集一样，符号注释的注释已经包含在这个数据集中。虽然RWTH-凤凰-Weater2014T和How2Sign提供了SLP评估基准，但它们不足以推广SLP模型。此外，这些数据集只包括德语和美国语的句子。为了为聋人和听力社区之间的相互交流提供一个易于使用的应用程序，需要新的大规模数据集，具有足够的多样性和不同的手语。关键是，这些符号通常是灵巧的，签名程序涉及到不同的渠道，同时包括手臂、手、身体、注视和面部表情。要捕捉这样的手势，需要在捕捉成本、测量（空间和时间）精度和生产的自发性之间进行权衡。此外，不同的设备用于数据记录，如有线电子爱、波磁传感器、配备红外摄像机的耳机、发射二极管和遥控器。上述设备捕获的不同通道之间的同步是数据收集和注释的关键。另一个挑战是使用一些捕捉设备，如捕捉手运动的复杂性，如电子爱。数据记录过程中的硬校准和偏差是这些采集设备的一些差异。外部设备的同步、手建模精度、数据丢失、捕获过程中的噪声、面部表情处理、注视方向和数据注释都是额外的挑战。考虑到这些挑战，为SLP提供一个大型和多样化的数据集，包括口语和手语注释，是非常不同的。图2显示了SLP的现有数据集。

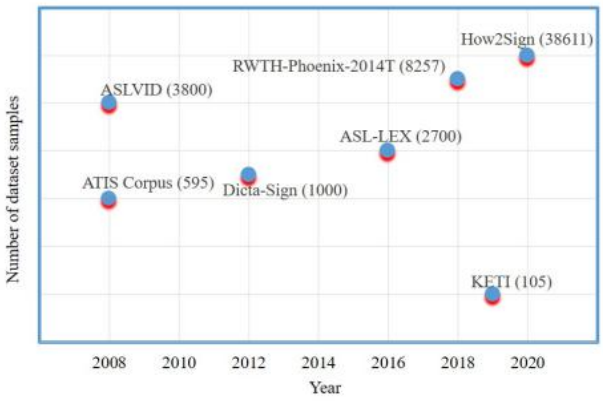

图2。SLP数据集的时间。每个数据集的样本数显示在括号中

2.3.应用程序

近年来，随着强有力的方法和技术的出现，机器翻译应用程序变得更加高效和值得信赖。机器翻译的早期努力之一可以追溯到60年代，在那里提出了一个从俄语翻译到英语的模型。该模型将机器翻译任务作为加密和解密的一个阶段。如今，标准的机器翻译模型主要可分为三类：基于规则的语法模型、统计模型和基于实例的模型。基于深度学习的模型，如Seq2Seq和NMT模型，属于第三类，并在SLP中显示出了有希望的结果。要将源语言翻译为目标语言，需要一个语料库来执行一些预处理步骤，包括边界检测、单词标记化和分块。虽然大多数口语都有不同的语料库，但手语缺乏如此庞大而多样化的语料库。美国手语(ASL)是世界上最大的手语社区，是为SLP开发的应用程序中使用最多的手语。由于聋人可能无法阅读或写口语，他们需要一些工具来与社会上的其他人交流。此外，许多有趣和有用的应用程序是不可访问的聋人社区。然而，我们还远没有为在现实场景中拥有大量词汇/句子的聋人提供应用程序。这些应用程序面临的主要挑战之一是使用许可权。只有一些应用程序是免费的。另一个挑战是缺乏对当前应用程序的泛化，这些应用程序是为非常特殊的应用程序场景的需求而开发的。

2.4.提出的模型

在本节中，我们将回顾在SLP中的最新工作。这些工作在ive类别中呈现和讨论：化身方法、NMT方法、运动图(MG)方法、条件图像/视频生成方法和其他方法。表1总结了SLP中回顾的模型。

2.4.1阿凡达方法

为了减少两者之间的沟通障碍
听力和听力受损的人，手语口译员被用作一种有效但代价昂贵的解决方案。为了在手头没有口译员的情况下迅速告知聋人，研究人员正在研究提供内容的新方法。其中一种方法是签名化身。《阿凡达是一种在没有与人类签名者对应的视频的情况下显示签名对话的技术。为此，我们采用了3D动画模型，与视频相比，它可以更有效地存储。手指、手、面部手势和身体的动作都可以使用化身来产生。这种技术可以被编程成用于不同的手语。近年来，随着计算机图形学的出现，计算机和智能手机可以生成高质量的动画，并在符号之间平稳过渡。为了捕捉聋人的运动数据，我们使用了一些特殊的摄像机和传感器。此外，还考虑一种计算方法将身体运动转移到符号化身[45]中。
两种获得符号化身的方法包括运动捕获数据和参数化光泽。近年来，一些作品已经开发出探索化身动画的参数化光泽。[7]，Tessa[18]，eSign[92]，符号[25]，赌博[34]，和WebSign[39]是其中的一些。这些工作需要通过转录语言注释的符号视频，如HamNoSys[63]或SigML[43]。尽管，这些化身的不受欢迎使他们不利于聋人社区。不清晰、不自然的动作，以及缺少非人工信息，如眼睛注视和面部表情，是化身方法的一些挑战。这些挑战导致了对国际手语序列的误解。此外，由于恐怖谷，用户感觉不舒服的[58]与机器人运动的化身。为了解决这些问题，最近的工作集中在注释非手动信息，如面部、身体和面部表情[23,24]。
使用从动作捕捉中收集的数据，角色可以更有用和接受(例如MocapLab[31]的Sign3D项目)。高度现实的结果是通过化身来实现的，但结果仅限于一小部分短语。这来自于数据的成本
收集和注释。此外，化身数据不是一个可伸缩的解决方案，需要专家知识来对生成的数据执行完整性检查。为了解决这些问题并提高性能，我们使用了基于深度学习的模型，作为最新的机器翻译开发项目。生成模型和一些图形技术，如运动图，最近被使用[79]。

2.4.2 NMT方法

机器翻译器是一种从一种语言翻译到另一种语言的实用方法。第一个翻译者回到了60年代，在那里俄语被翻译成英语的[38]。翻译任务需要对源语言进行预处理，包括句子边界检测、单词标记化和分块。这些预处理任务具有挑战性，特别是在手语方面。手语翻译(SLT)的目的是从不同的词序和语法中产生/生成口语翻译。注释的顺序和数量不需要与口语句子中的单词相匹配。目前，有不同类型的机器翻译器，主要基于语法规则、统计数据和例子[60]。作为一个基于例子的方法，一些研究工作已经发展到利用人工神经网络(ann)从文本翻译成手语，即NMT[6]。NMT使用人工神经网络来预测一个单词序列的可能性，通常在一个集成模型中建模整个句子。为了提高长序列的翻译性能，Bahdanau等人。[6]提出了一种有效的注意机制。这一机制后来被Luong等人改进。
[51].Camgoz等。提出了一种将seq2seq模型与CNN的结合，将符号视频翻译为口语句子[12]。郭等人。[35]设计了一个混合模型，包括结合3D卷积神经网络(3DCNN)和基于长短期记忆(LSTM)的[56,37]编码器-解码器，从符号视频转换为文本输出。.071在他们自己的数据集上的结果显示，与最先进的模型相比，精度度量提高了0%。扩张卷积和变换器也是用于手语翻译[40,86]的两种方法。斯托尔等人。[79]提出了一种使用NMT、GANs和运动生成的自动SLP混合模型。该模型从口语句子中生成符号视频，并将最少的数据注释用于训练。这个模型首先将口语句子翻译成符号姿态序列。然后，利用生成模型生成可信的手语视频序列。在PHOENIX14T手语翻译数据集上的结果显示了可比性的结果

适合使用最先进的替代品。
虽然基于NMT的方法在翻译任务中取得了成功，但还需要解决一些重大挑战。领域适应是这一领域的第一个挑战。由于不同域之间的翻译受到不同规则的影响，因此领域自适应是开发针对特定用例的机器翻译系统的关键要求。第二个挑战是关于训练数据的数量。特别是在基于深度学习的模型中，增加数据量可以带来更好的结果。另一个不同的是处理不寻常的词语。翻译模型在这些词上表现很差。单词对齐和调整波束搜索参数是基于NMT的模型面临的其他挑战。目前基于深度学习的模型的良好结果为这一领域的未来研究奠定了基础。

.4.32运动图方法

运动图(MG)是一种动态动画人物的计算机图形方法，被定义为由运动捕捉数据构造的有向图。MG可以生成新的序列来满足特定的目标。在SLP中，MG可以与基于NMT的网络相结合，进行连续的文本到姿态的翻译。对MG的早期研究之一可以追溯到2002年，当时由Kovar等人提出了一个总体框架。[47]，用于提取满足用户专长的特定图形行走。两帧之间的距离被定义为两点云之间的距离。为了进行过渡，在关节之间使用了运动和位置的对齐和插值。最后，将分支搜索算法和边界搜索算法应用于该图。在另一项工作中，Arikan和Forsyth[1]使用联合位置、速度和加速度参数来确定两个连续帧之间的距离。此外，使用平滑函数计算了两个剪辑之间的不连续性。总结图后，对图进行随机搜索。Lee等人提出了另一种方法中的运动数据的两层表示。[49].在第一层，数据被建模为一阶马尔可夫过程，并利用加权关节角和速度的距离计算转移概率。对第二层，即聚类林进行了聚类分析，以推广运动。斯托尔等人。[79]提出了一种利用姿态数据的连续SLP的MG。符号光泽被嵌入到MG中，每个时间步由NMT解码器提供的转移概率。
虽然MG可以通过运动捕捉数据库生成可信和可控的运动，但它也面临着一些挑战。第一个挑战是关于对数据的访问。为了用真正多样化的动作来显示模型的潜力，需要大量的数据。可扩展性和com-
选择最佳过渡的图形的计算复杂性是MG中的其他挑战。此外，由于离开节点的边数随着图的大小的增加而增加，因此搜索算法中的分支因子也会增加。

.4.42有条件的图像/视频生成

近年来，自动图像/视频生成经历了显著的发展。然而，视频生成的任务具有挑战性，因为结果帧之间的内容必须是一致的，显示一个合理的运动。由于对人类视频生成的需要，这些挑战在SLP中更加不同。这些视频中的动作和出现的复杂性和多样性是很高的和具有挑战性的。控制生成视频的内容是至关重要的，但又不同。
随着深度学习的最新进展，自动图像/视频生成已经出现了使用基于神经网络的架构的不同方法，如CNNs[16,84]、RNNs[33,83]、变分自动编码器(VAEs)[44]、条件VAEs[89]和GAN[32]。VAEs和GANs通常结合在一起，以有利于VAE的稳定性和GAN的鉴别性质。与SLP最相关的是一种混合模型，包括VAE和GAN组合，已被提出用于人[52,77]的图像生成和执行手语[78,85]的人的视频生成。此外，还有一些用于图像/视频生成的模型可以用于SLP。例如，Chen和Koltun[16]提出了一个基于cnn的模型来生成给定语义标签地图的摄影图像。范登德等人。[84]提出了一种基于深度学习的模型，即Pixelrns，沿二维依次生成图像像素。格雷戈尔等人。[33]开发了一种基于RNN的架构，包括一个编码器和一个解码器网络来压缩训练过程中呈现的真实图像和接收代码后的正弦图像。卡拉斯等人。[41]设计了一个名为StyleGAN的深度生成模型，来调整每个卷积层的图像风格。Kataoka等。[42]提出了一个结合了GAN和注意机制的模型。由于来自注意机制，该模型可以生成包含高详细内容的图像。虽然基于深度学习的生成模型最近取得了显著的成果，但在其训练中仍存在着重大挑战。模态崩溃、不收敛和不稳定、合适的目标函数和优化算法都是这些挑战。然而，最近提出了一些策略来更好地设计和优化它们。适当的网络架构设计、适当的目标函数和优化算法是旨在提高基于深度学习的模型性能的一些技术之一。

2.4.5其他模型

除了之前的类别外，我们还提出了一些使用不同的深度学习模型的SLP模型。例如，桑德斯等人。[73]提出了一种渐进变形器，作为一种基于深度学习的模型，从口语句子中生成连续的符号序列。在另一项工作中，Zelinka和Kanis[91]设计了一个专注于骨骼模型生产的手语合成系统。一个前馈变压器和一个循环变压器，作为基于深度学习的模型，以及注意机制已经被用来提高模型的性能。桑德斯等人。[75]提出了一种基于生成的模型，从文本输入中生成逼真的连续符号视频。他们结合了一个转换器和一个混合密度网络(MDN)来管理从文本到骨骼姿势的转换。托内等人。[81]设计了一种使用多通道信息（手的形状、手的运动、鼠标移动、面部表情）的SLP评估方法。在这种方法中，考虑了两个语言方面：生成的词汇素和生成的形式。在这类方法中使用不同方法的功能已经导致了SLP的成功结果。然而，模型复杂性的挑战仍然是一个有待解决的问题。在准确性和准确性之间做出权衡。任务的复杂性是一个关键因素。

3.讨论

在这项调查中，我们详细回顾了SLP的最新进展。我们提出了一个分类法，它总结了与SLP相关的主要概念。我们对SLP中最近的工作进行了分类，并在每个类别中提供了单独的讨论。所提出的分类法涵盖了不同的输入模式、数据集、应用程序和所提出的模型。在这里，我们总结了主要的信息：

输入模式：通常，视觉模式和语言模式是SLP中的两种输入模式。视觉模态包括捕获的图像/视频数据，而语言模态包含来自自然语言输入的文本。计算机视觉和自然语言处理技术是处理这些输入模式的必要条件。这两类方法都受益于深度学习方法来提高模型性能。RGB和骨架是SLP模型中使用的两种常见的视觉输入数据类型。RGB图像/视频包含高分辨率内容，骨架输入降低了模型的参数复杂性，有助于制作低复杂和快速的模型。GAN和LSTM是SLP中最常用于视觉输入的基于深度学习的模型。虽然使用这些模型取得了成功的结果，但需要更多的努力来生成更逼真和高分辨率的符号图像/视频。在基于深度学习的语言模态模型中，NMT模型是输入文本处理中最常用的模型。其他的Seq2Seq模型，如基于RNN的模型，在许多任务中证明了它们的有效性。虽然使用这些模型获得了准确的结果，但需要更多的努力来克服翻译任务中现有的挑战，如领域适应、不常见的单词、单词对齐和单词标记化。

数据集：缺乏大型注释数据集是SLP面临的主要挑战之一。手语数据的收集和注释是一项昂贵的任务，需要语言专家和母语者的合作。虽然有一些公开可用的SLP[4,9,14,15,22,46,54]数据集，但它们的手语注释数据却很薄弱。此外，SLP中的大多数可用数据集包含vocabb-的限制域laries/sentences.为了在聋人和听力社区之间进行现实世界的交流，有必要获得一个在句子层面上分割的大规模连续手语数据集。在这样的数据集中，需要包含一个连续的手语句子和相应的口语句子的成对形式。只有少数数据集符合这些标准，[12,22,46,91]。关键是，上述的大多数数据集不能用于端到端翻译[12,46,91]。两个公共数据集，RWTH-凤凰城-2014T和How2Sign，是SLP中使用最多的数据集。前者包括Ger可用于文本到手语翻译的人类手语句子。后者是最近提出的一个用于语音到符号的语言翻译的多模态数据集。虽然RWTH-凤凰城-天气2014T[12]和How2Sign[21]提供了适当的SLP评估基准，但它们还不足以推广SLP模型。此外，这些数据集只包括德语和美国语的句子。从口语翻译成大量多样化的手语是聋人社区面临的一个主要挑战。

应用程序：美国手语(ASL)是世界上最大的手语社区，是为SLP开发的应用程序中最常用的手语。由于聋人可能很难读到或写出口语，所以他们需要一些工具来与社会上的其他人进行交流。此外，许多有趣和有用的应用程序是不可访问的聋人社区。为了解决这些挑战，已经提出了一些旨在开发这些工具的项目。虽然这些应用程序成功地在聋人和听力社区之间架起了桥梁，但我们还远没有涉及来自复杂现实场景的大型词汇/句子的应用程序。这些应用程序面临的主要挑战之一是使用许可权。另一个挑战是关于应用程序领域。这些应用程序大多数都是为非常特殊的领域开发的，如诊所、医院和po-

表1。深度SLP模型的总结。

年	裁判员	特征	输入方式	数据集	描述
2011	[45]	化身	RGB视频	ViSiCAST	优点。提出了一个基于光泽的工具，重点关注动画内容的评估，使用一个新的度量来比较化身与人类签名者。缺点。需要包括人工签名者的非手动功能。
2016	[55]	化身	RGB视频	自己的数据集	优点。自动添加生成图像的真实性，低计算复杂度。缺点。需要将肩膀和躯干伸展的位置放在化身的肘部的位置上，而不是IK末端执行器。
2016	[31]	化身	RGB视频	自己的数据集	优点。易于理解与高观众接受的标志化身的缺点。仅限于少量的符号短语
2018	[12]	nmt	RGB视频	凤凰城 天气 2014T	优点。强大的联合对齐、识别和翻译符号视频。缺点。需要在空间域中排列这些符号。
2018	[35]	nmt	RGB视频	自己的数据集	优点。与句子中视觉内容对应的词序对齐。缺点。需要泛化到其他数据集。
2020	[79]	NMT, MG	文本	凤凰城14吨	优点。健壮到最小的光泽度和骨骼级注释的模型训练。缺点。模型复杂度高。
2020	[73]	其他	文本	凤凰城14	优点。对输出符号序列的动态长度具有鲁棒性。缺点。模型的性能可以得到提高，包括非手动信息。
2020	[91]	其他	文本	捷克新闻	优点。坚固的骨骼部分。缺点。模型的性能可以得到提高，包括面部表情的信息。
2020	[75]	其他	文本	凤凰城14吨	优点。强大的非手动功能生产。缺点。需要增加所生成的符号的真实性。
2020	[13]	其他	文本	凤凰城14吨	优点。不需要有光泽度的信息。缺点。模型复杂度高
2020	[74]	其他	文本	凤凰城14吨	优点。强大的手动功能生产。缺点。需要增加所生成的符号的真实性。

虱子站。提高可用数据的数量及其质量将有助于创建这些所需的应用程序。

提出的模型：在SLP中提出的工作可以分为类别：阿凡达方法、NMT方法、MG方法、条件图像/视频生成方法和其他方法。表1显示了最先进的深度SLP模型的摘要。生成的视频和注释注释的一些示例如图3、4和5所示。使用从动作捕捉中收集的数据，角色可以更有用和可接受。化身获得了高度现实的结果，但结果仅限于一小的短语。这来自于数据收集和注释的成本。此外，化身数据不是一个可伸缩的解决方案，需要专家知识来检查和抛光。为了解决这些问题和提高性能，我们使用了基于深度学习的模型。

虽然基于NMT的方法在翻译任务中取得了显著的效果，但还需要解决一些重大挑战。

领域适应是这一领域的第一个挑战。由于不同领域的翻译需要不同的风格和需求，因此这是开发针对特定用例的机器翻译系统的关键要求。第二个挑战是关于可用的培训数据的数量。特别是在基于深度学习的模型中，增加数据量可以带来更好的结果。另一个挑战是关于这些不寻常的词语。翻译模型在这些词上表现很差。单词对齐和调整波束搜索参数是基于NMT的模型中面临的其他挑战。

虽然MG可以通过运动捕捉数据库生成可信和可控的运动，但它也面临着一些挑战。第一个挑战是关于对数据的有限访问。为了用真正多样化的动作来显示模型的潜力，需要大量的数据。通过对图的可扩展性和计算复杂性来选择最佳过渡是MG中的其他挑战。此外，由于离开节点的边数随着图大小的增加而增加，搜索算法中的分支因子会

也会增加。为了自动调整图的一致性，并依赖于训练数据，可以使用图卷积网络(GCN)和一些控制算法，将图的结构单调地采用，而不是用户干扰。

虽然GANs最近在图像/视频生成方面取得了显著的成果，但在GANs的训练方面存在着重大挑战。模态崩溃、不收敛和不稳定、合适的目标函数和优化算法都是这些挑战。然而，最近有人提出了一些建议来解决更好的设计和优化问题。适当的网络架构设计、适当的目标函数和优化算法是被提出的用来提高基于gan的模型性能的一些技术。最后，混合模型的模型复杂性仍然面临着挑战。

局限性：在本调查中，我们介绍了使用深度学习在SLP和相关领域的最新进展。虽然最近的基于深度学习的模型已经在SLP中取得了成功的结果，但仍有一些局限性需要解决。一个主要的挑战是关于多重签名者(MS)的一代，这是在聋人社区中提供现实世界的交流所必需的。为此，我们需要产生多个不同外观和一致性的签名者。另一个限制是高分辨率和逼真的连续手语视频的可能性。在SLP中提出的大多数模型只能生成低分辨率的符号样本。对从训练数据中提取的人工关键点进行调节，可以降低模型的参数复杂度，并有助于产生高分辨率的视频符号。然而，基于虚拟角色的模型可以成功地生成高分辨率的视频样本，尽管它们非常复杂且昂贵。此外，MG的剪枝算法还需要改进，包括手语的额外特征，如持续时间和速度。

未来的发展方向：虽然最近的SLP模型依靠深度学习能力提供了有希望的结果，但仍有很大的改进空间。考虑到自我注意的辨别能力，学习融合多种输入模式以适应多通道信息，学习结构化的时空模式（如图神经网络模型），以及使用领域专业先验知识是该领域未来的发展方向。此外，还有一些令人兴奋的辅助技术，为聋人和听力受损者提供帮助。简要介绍这些技术可以深入了解SLP领域的研究人员，并在它们与相应的技术需求之间架起桥梁。这些技术可分为三类设备：听力技术、警报设备和通信支持技术。例如，让imag-

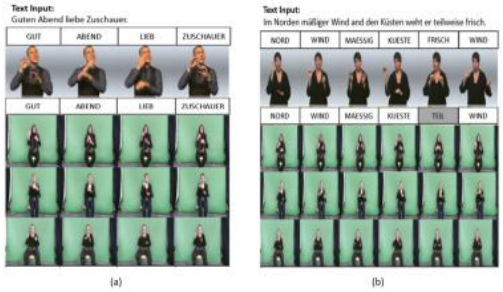

图3。翻译结果来自[79]：(a)“(a)”。（晚上好，亲爱的观众们），(b)，“我是诺顿小姐小姐”。（北方有微风，海岸有新鲜吹来）。上面一行：地面真实光泽和视频，下面一行：生成光泽和视频。该模型结合了NMT网络和SLP的GAN。

图4。来自[79]的翻译结果：来自口语的文本被翻译成人类的姿势序列。

图5。来自[45]的翻译结果：使用角色动画系统创建了一个签名化身。上一行：签名头像，下一行：原创视频。

这是一种帮助聋人将音乐体验转化为另一种感官模式的技术。虽然SLP的最新进展是有希望的，但在考虑快速手运动的不受控制的环境中提供一个快速处理模型是不可必要的。很明显，技术标准化和设备和平台之间的完全互操作性是在听力和听力受损社区之间进行现实生活中的交流的先决条件。

这项工作得到了HIS公司和伊朗基础科学研究所(IPM)、西班牙项目PID2019-105093GB-I00(MINECO/Feder，UE)、加泰罗尼亚CERCA项目和ICREA学术项目的部分支持。

参考文献

[1]OkanArikan和D.A.。福赛思盖尔语人名的英语形式从示例中进行交互式运动生成。*在第29届计算机图形和交互技术年会的会议记录中，签名图‘02，第483-490页，2002年*。

[2]启动美国手语。中文手语。*https://www.startasl.com/中文手语*, 2021.

[3]启动美国手语。西班牙语手语。*https://www.startasl.com/西班牙语手语ss1/*, 2021.

[4]瓦西利斯阿西索斯，卡罗尔尼德斯，斯坦斯克拉夫，琼纳什，亚历山德拉斯特凡，全袁，和阿什温坦加利。美国手语词汇的视频数据集。CVPR，2018年第1-8页。

[5]丹尼尔·巴赫曼，弗兰克·魏克特和格哈德·林克诺尔。以跳跃动作控制器为重点的三维人机交互的回顾。*传感器，2018年*。

[6]，赵庆贤和本国。神经机器翻译通过联合学习来对齐和翻译。ICLR, 2015.

[7] J.A.Bangham, S.J.考克斯。艾略特，J。R.W.格劳尔特，我。马歇尔，S。兰科夫和M。泉虚拟签名：捕获、动画、存储和传输-视觉项目的概述。*针对残疾人和老年人的语音和语言处理，2000年*。

[8]，马克·博格和肯尼斯·P。卡米莱里。基于深度学习的手语识别的语音和有意义的子单元。ECCV, 2020.

[9]Jan邦格罗斯，丹尼尔·斯坦，菲利普·德鲁，赫尔曼·内，萨拉·莫里西，安迪·韦和莱内特·范·齐尔。有了一个自由的手语语料库。*第六届国际语言资源与评价会议，2008年*。

[10] A.H.屁股，E。Rovini, C.Dolciotti, G.德·彼得里斯，P。蒙吉奥尼，M。C.卡博尼和F。卡瓦略利用跳跃运动控制器对帕金森病的目的和自动分类。*生物医学在线，2018年17日*。

[11]苏蔡、朱高霞、吴英天、刘英瑞、胡孝义。一个基于手势的游戏在提高基本运动技能和识别能力方面的案例研究。*交互式学习环境，2018年26日*。

[12]NecatiCihanCamgoz，西蒙·哈迪尔德，奥斯卡·科勒，赫尔曼·尼伊和理查德·鲍登。神经手语翻译。CVPR, 2018.

[13]NecatiCihan坎戈兹，奥斯卡·科勒，西蒙·哈迪尔德和理查德·鲍登。用于多发音手语翻译的多通道变压器。*ECCVW, 2020*.

[14]NecatiCihanCamgz，西蒙·哈迪尔德，奥斯卡·科勒，赫尔曼·尼伊和理查德·鲍登。t：手语视频、光泽和翻译的平行语料库。*CVPR，盐湖城，UT，2018年*。

[15]娜奥米K。卡塞利，泽德塞夫奇科娃塞希尔，阿里尔M。科亨戈德堡和凯伦·埃莫瑞。一个针对asl的词汇数据库。行为研究方法，49：784-801,2017。

陈[16]峰和科尔敦。利用级联恢复网络进行的摄影图像合成。ICCV，第1511-1520页。

[17]Miri维斯科恩，以色列沃尔德曼，丹尼尔雷加佐尼，和安德里亚维塔利。通过手势识别进行手部康复使用跳跃运动控制器。*第11届人类系统交互国际会议(HSI)，波兰格但斯克，2018年*。

[18]斯蒂芬考克斯，迈克林肯，朱迪特雷格瓦森，梅勒妮中莎，马克威尔斯，马库斯塔特，和桑贾阿伯特。这，一个帮助与聋人交流的系统。*在第五届国际ACM关于辅助技术的会议的论文集上，第205-212页，2002年*。

[19]哈立德。达拉布克，法拉H。阿尔图克和萨达兹。斯威丹。带有文本聊天系统的三维虚拟现实合作绘画教育应用。Comput Appl Eng Educ, 26:1677 - 1698, 2018.

[20]Fel克斯道斯，杰奎斯渲染，朱塞佩卡波恩。使用跳跃传感器远程控制机械手。*IFToMM意大利国际会议，第332-341页，2018年*。

[21]阿曼达杜阿尔特。跨模态的神经手语翻译。*第27届国际会议ACM*。

[22]阿曼达·杜阿尔特，什鲁蒂·帕拉斯卡尔、迪普蒂·加迪亚拉姆、肯尼斯·德汉、弗洛里安·梅策、乔迪·托雷斯和泽维尔·吉罗伊涅托1。如何2符号：一个用于连续美国手语的大规模多模态数据集。*手语识别、翻译和制作研讨会，2020年*。

[23]莎拉·埃布林斯和马特·赫纳福斯。利用序列分类来弥合手语机翻译和手语动画之间的差距。*SLPAT 2015年的论文集：第六届辅助技术的语音和语言处理研讨会，第2-9页，2015年*。

[24]莎拉说约翰和格劳特约翰·格劳特。充分利用茉莉花的潜力来建立一个签署火车公告的化身。*第三届手语翻译与化身技术国际研讨会，2013年第1-9页*。

[25]Eleni，斯塔夫鲁拉-埃维塔·福蒂娜，托马斯·汉克，约翰·格劳特，理查德·鲍登，布拉福特，克里斯托夫·科莱特，马拉戈斯和弗朗克·列斐伏尔-阿尔巴雷特。格言签名维基：为聋人提供网络交流。*国际残疾人计算机会议(ICCHP)，第205-212页，2012页*。

[26]民族学。阿根廷的手语。*https://www.民族学。com/language/aed*, 2021.

[27]民族学。希腊语的手语。*https://www.民族学。com/language/gss*, 2021.

[28]民族学。波斯语的手语。*https://www.民族学。com/language/psc*, 2021.

[29]延斯·福斯特，克里斯托夫·施密特，托马斯·霍尤克斯，奥斯卡·科勒，乌维·泽尔，贾斯图斯·皮亚特和赫尔曼·内。天气：大词汇手语识别和翻译语料库。*第八届语言资源与评价国际会议记录(LREC12)，土耳其伊斯坦布尔，第3785-3789页，2012年*。

[30]SakherGhanem，克里斯托弗·科利，和瓦西利斯·阿西索斯。一项关于使用智能手机进行手语识别的调查。*第十届与辅助环境相关的普及技术国际会议记录，希腊罗德岛，2017年*。

[31]西尔维·吉贝特，阿尔巴雷特，卢多维奇·哈蒙，瑞米·布伦和艾哈迈德·图尔基。法语交互式编辑

专门针对虚拟签名者的手语：需求和挑战。*信息社会的普遍获取，15：525-539,2016。*
[32]伊恩J。古德费罗、阿巴迪、迈赫迪米尔扎、徐平、法利、大卫、奥泽尔、库维尔和本吉奥。生成的对抗网络。*神经信息处理系统的研究进展，第2672-2680页，2014年。*
[33]卡罗尔·格雷戈尔，伊沃·丹尼赫尔卡，亚历克斯·格雷夫斯，丹尼洛·雷森德，和达恩·维尔斯特拉。绘制：一种用于图像生成的递归神经网络。*《机器学习研究论文集》，2015年。*
[34]虚拟人类组。针对手语动画的虚拟人类研究。*计算科学学院，UEA诺里奇，英国，2017年。*
[35]郭丹，周文刚、李后强、王孟。手语翻译的分层lstm。*第三十届AAAI艺术智能会议(AAAI18)，2018年。*
[36]托马斯汉克。德语手语。
https://www.awhamburg.长期研究scientihc-projects/dictionary-german-sign-language.html, 2021.
[37]九月假日和施米杜伯。长期记忆。*神经计算，1997年9日。*
[38]约翰哈钦斯。机器翻译的历史。
*http://psychotransling.ucoz.com/ld/0/11-哈钦斯调查*pdf, 2005.
[39]穆罕默德·杰姆尼，食尸鬼，梅雷兹·布拉尔，努本·亚希亚，贾巴拉，阿克拉夫·奥斯曼，和尤内布。网络标志。*http://www.latice.rnu.tn/websign/, 2020.*
[40]纳尔·卡尔奇布伦纳，拉斯·埃斯佩霍尔特，卡伦·西蒙扬，亚伦·范登奥德，亚历克斯·格雷夫斯和科雷·卡沃库格鲁。线性时间内的神经机器平移。*arXiv:1610.10099, 2016.*
[41]TeroKarras，萨穆利·莱恩，和蒂莫·艾拉。一种基于风格的生成式对抗网络的生成器架构。CVPR, 2019.
片冈[42]，松原隆和上原久明。使用对抗性网络和注意机制生成图像。*IEEE/第15届计算机与信息科学国际会议(ICIS)，2016年。*
[43]理查德·肯纳威。为实时手势动画的化身独立的脚本。*手语的程序性动画，arXiv：1502.02961，2013。*
[44]饮食协会，金玛和马克斯·韦林。自动编码变分贝叶斯。ICLR, 2014.
[45]迈克尔·基普，亚历克西斯·赫洛伊尔和阮全。手语化身：动画和可理解性。*智能虚拟代理国际研讨会，第113-126页，2011年。*
高[46]，金张乔，郑家东，赵中尚。基于人的关键点估计的神经手语翻译。*应用科学，2019年9日。*
[47]·卢卡斯·科瓦尔，迈克尔·格莱彻和弗雷德里克·皮金。运动图。*签名图‘02：第29届计算机图形和交互技术年会的会议记录，第473-483页，2002年。*
[48]Agelos克拉提梅诺斯，乔治斯帕夫拉科斯和马拉戈斯。3d的手，面部和身体提取的手语识别。ECCV, 2020.
李杰熙和宋勇信。一种针对类似人的交互式动作编辑的分级方法。*签名图‘99：第26届计算机图形和交互技术年会的会议记录，第39-48页，1999年。*
[50]XingLiang、安杰斯塔尼奥斯、本西·沃尔、巴塔特和泰伦·伍尔夫。多模态机器学习方法和工具包，自动识别痴呆的早期阶段。ECCV, 2020.
[51]明-阮，何潘和克里斯托弗·D。人员配备基于注意力的神经机器翻译的有效方法。*arXiv:1508.04025, 2015.*
马[52]、徐佳、孙千如、席勒、图恩、古乐。姿态引导下的人物形象生成。nip，2017年。
[53]Nezam马吉迪，库罗什基亚尼，和拉齐赫拉斯古。一个从单一图像中进行超分辨率增强的深度模型。人工智能和数据挖掘杂志，8：451-460,2020。
[54]西尔克马特斯，托马斯汉克，安贾雷根，雅各布斯托兹，萨图沃塞克，埃莱尼埃夫西米乌，亚塔纳西亚-丽达迪莫，安妮丝布拉福特，约翰格劳特，和伊娃萨法尔。构建一个多语言的手语语料库。*在第五LREC。伊斯坦布尔，2012年。*
[55]：约翰·麦克唐纳、罗莎莉·沃尔夫、杰里·施奈普、朱莉·霍奇格桑、戴安娜·戈尔曼·贾姆罗齐克、玛丽·斯顿博、拉尔万·伯克、梅丽莎·比亚莱克和法拉·托马斯。一种实时制作逼真的美国手语动画的自动技术。信息社会的普遍获取，15：551-566,2016。
[56]长短期内存。9月和我。*神经计算，1997。*
[57]Matteo·莫兰多，塞雷娜·庞特，伊利莎·费拉拉，和西尔瓦娜·德勒皮恩。通过严肃的游戏来确定家庭康复的运动和生物物理指标。*信息，2018年9日。*
森[58]，卡尔F。麦克多曼和诺里·卡吉基。来自爱尔兰的神秘谷。*IEEE《机器人与自动化》杂志，19：98-100,2012。*
[59]阿米特莫约塞夫，约约尼斯，罗伊阿哈罗尼，萨拉埃布林，和斯里尼纳拉亚南。使用人的姿态估计的实时手语检测。ECCV, 2020.
[60]AchrafOthman和默罕默德·杰姆尼。统计手语机翻译：从英语文字到美国手语光泽。*IJCSI国际计算机科学杂志，8：65-73,2011。*
[61] Owlcation.韩语的手语。*https://owlcation.韩语手语,* 2021.
[62]玛丽亚·帕雷利，卡特琳娜·帕帕迪米特里奥，杰拉西莫斯·波塔米亚诺斯，乔治奥斯·帕夫拉科斯和彼得罗斯·马拉戈斯。利用rgb视频中基于深度学习的手语识别中的三维手姿态估计。ECCV, 2020.

[63] Siegmund Prillwitz.哈姆诺西斯。版本2.0。汉堡手语符号系统。介绍性指南。*汉堡信号出版社，1989年。*
[64]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。利用限制性玻尔兹曼机对静止图像进行多模态深度手符号识别。*熵，2018年20日。*
[65]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。使用多视图手骨架进行手手语识别。*具有应用程序的专家系统，150,2020年。*
[66]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。基于视频的深度级联模型分离手手语识别。多媒体工具和应用程序，79：22965-22987,2020。
[67]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。手姿态感知的多模态孤立的手语识别。*多媒体工具和应用程序，80：127-163,2021。*
[68]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。使用深度网络和svd实时隔离手手语识别。*《环境智能与人源化计算杂志》，2021年。*
[69]Razieh•拉斯特古，库罗什•基亚尼和塞尔吉奥•埃斯卡莱拉。手语识别：一项深入的调查。*带有应用程序的专家系统，164：113794,2021。*
[70]马可•罗切蒂，古斯塔沃•玛丽亚和安吉洛•塞梅拉罗。野外游戏：在公共空间游戏的基于手势的界面。*视觉传达与图像表现杂志，23：426-440,2012。*
[71]萨尔托。波兰语的手语。*https://www.萨尔托青年。网络/工具/otlas合作伙伴hnding/organisation/association-of-polish-sign-languageinterpreters.2561/，2021.*
[72]本•桑德斯，坎戈兹和理查德•鲍登。多渠道手语制作的对抗性培训。BMVC，2020.
[73]本•桑德斯，坎戈兹和理查德•鲍登。用于端到端手语制作的渐进式变形金刚。ECCV，第687-705页，2020年。
[74]本•桑德斯，Camgz和理查德•鲍登。多渠道手语制作的对抗性培训。BMVC，2020.
[75]本•桑德斯，Camgz和理查德•鲍登。现在大家都签名：将口语翻译成照片逼真的手语视频。*arXiv:2011.09846,2020.*
[76]阿比盖尔看到和马修•拉姆。机器翻译，序列到序列和注意。*https://web.斯坦福。edu/class/cs224n/slides/cs224n-2020-lecture08-nmt.*pdf，2021.
[77]，西亚罗欣，桑吉尼托，斯蒂芬•拉图利埃尔，和尼古•塞贝。可变形的通用程序算法，用于基于姿态的人体图像生成。CVPR，2018.
[78]斯蒂芬妮•斯托尔，内卡蒂•西汉•坎戈兹，西蒙•哈迪尔德和理查德•鲍登。使用神经机器翻译和生成式对抗网络制作手语。*BMVC，2018.*
[79]斯蒂芬妮•斯托尔，内卡蒂•西汉•坎戈兹，西蒙•哈迪尔德和理查德•鲍登。为了使用神经机器翻译和生成式加法器来制作手语萨里亚网络。*国际计算机视觉杂志，128：891-908,2020*。
[80]Ilya苏茨克弗，葡萄酒和报价V。黎巴嫩用神经网络进行序列学习。nip，2014年。
[81]桑德林龟，花龟，理查德鲍登和麦吉麦多斯。一种基于音位学的手语孤立符号产生评估方法。*ICMI‘20同伴：2020年多模式交互国际会议的同伴出版，2020年。*
[82]奥雷里尤斯•维特克维乌斯，曼塔斯•塔罗萨，托马斯•布劳斯卡斯，罗伯塔斯•达马维乌斯，里蒂斯•马斯克利纳斯和马辛•沃尼亚克。在虚拟现实中使用跳跃动作来识别美国手语的手势。*应用程序。科学。*，9，2019.
[83]亚伦•范登Oord，纳尔•卡尔奇布伦纳和科雷•卡武库格鲁。像素递归神经网络。*第33届机器学习国际会议论文集，第1747-1756,2016页。*
[84]亚伦•范登奥德，纳尔•卡尔奇布伦纳，奥里奥尔维亚尔，拉斯•埃斯佩霍尔特，亚历克斯•格雷夫斯和科雷•卡夫科格鲁。使用像素网络解码器的条件图像生成。nip，2016年。
[85]尼尔•瓦萨尼，普拉蒂克•奥蒂安和萨米普•卡利亚尼；鲁希娜卡拉尼。通过句子处理和生成式对抗网络生成印度手语。*智能可持续发展系统国际会议(ICISS)，2020年。*
[86]什瓦斯瓦尼，诺姆沙泽，尼基帕马，雅各布乌斯科利特，狮子琼斯，艾丹N。戈麦斯，尤卡斯•凯泽，和伊利亚•波罗苏欣。你所需要的就是注意力。nip，2017年。
[87]威廉G。牧师。美国手语。*http://www.lifeprint.com/*，2021.
[88]世卫组织：世界卫生组织。耳聋和听力损失。*http://www.谁int/mediacentre/factsheets/fs300/en/，2021.*
[89]新臣、杨姬、宋喜贤、李宏乐。属性2图像：从视觉属性生成的条件图像。ECCV，第776-791页，2016年。
[90]Hee-Deok杨。基于条件随机序列的运动传感器的手语识别。*传感器，2014年15：135-147。*
[91]Jan泽林卡和JakubKanis。神经符号语言的合成：单词是我们的注释。WACV，第3395-3403页，2020年。
[92]Inge[92]，弗林登，约翰罗斯，和桑尼范德选择。给聋人的合成签名：设计。*http://www.visicast.cmp.uea.ac.uk*，第1-6页，2005年。